

Wanted: a voice for the life sciences

Academic life scientists are currently well-funded, but they need to develop better advocacy skills in order to exert some influence over issues related to their work.

These are among the best of times for many in the life sciences, with strong government and industrial support helping to develop a burgeoning community which makes an important contribution to both scientific discovery and public life. During such times, it makes sense for the community to ensure that it speaks with a clear voice, not just in the corridors of power, but also in our broader society, where so many voices are striving to be heard. So it is unsurprising that biologists in both the United States and the United Kingdom (see pages 310, 315, this issue) are striving for more effective representation from their professional societies and related advocacy groups.

A similar set of issues concern these scientists on both sides of the Atlantic. Close to the top of their agenda is graduate education. There is a growing realization within the community that the attraction of young people into a lengthy period of very hard work for very little money, to be followed in most cases by gradual disillusionment as the prospect of a tenure-track position fades, is a sub-optimal approach to career development. Increasingly pervasive restrictions on the use of research tools and research data make up another area in which academics need to speak up, before their voice is overwhelmed by industrial interests. Scientific integrity, animal care, cloning and genetic engineering are all issues in which the community can help itself and the public by communicating a coherent and credible point of view.

In the United States, the Federation of American Societies for Experimental Biology (FASEB), which publicly represents 17 scientific societies, has emerged in recent years as an influential voice in public policy. But it competes for attention with rivals such as the 40,000-strong American Society for Microbiology and the American Institute of Biological Sciences, a federation representing whole-organism biologists.

This fragmentation is not ideal, but FASEB has nonetheless proven its influence, especially in campaigning for money for the National Institutes of Health. It makes sense that it should seek to extend this success into other spheres. In Washington, it is particularly important that the academic community make itself heard: unlike the pharmaceutical and biotechnology industries, academics cannot afford to peddle influence by paying large sums of money to sympathetic politicians.

In the United Kingdom, a belief that scientific societies are failing to adequately represent life scientists has led to the foundation of a UK Life Sciences Committee, comprising representatives of 12 learned societies in the biosciences, which organized a meeting in London earlier this month on graduate education. The committee has identified a requirement — which it may itself duly fill — for a more broadly based organization to represent both academic and industrial science. It remains to be seen if the different interests of the two can be adequately reconciled.

Scientific societies with a longer tradition of high-level involvement in policy matters — such as the American Physical Society — have refined this craft to the point where their membership is clearly capable of punching its weight. The picture in biology is more complicated, with long-standing enmities (some of which can be traced to the sudden ascendancy of molecular biology in the 1960s, and the subsequent transformation of the field) serving to inhibit effective public advocacy. Umbrella groups such as FASEB and the UK Life Sciences Committee are constrained in what they can say by their need to consult with so many affiliates. The time is right for biologists to find a voice commensurate with their importance in society, but this may prove to be easier said than done. □

“Red-Green” warning signals

A change in Germany's government could place some sensitive science issues on centre stage.

In the campaign for this Sunday's federal election in Germany, the two main parties, Chancellor Helmut Kohl's Christian Democrats (CDU) and Gerhard Schröder's Social Democrats (SDP), have unsurprisingly chosen to focus their campaigns on unemployment, immigration and crime. The science policies of these parties are largely identical, and science issues will only come to the fore if the Green party forms part of the next government — the most likely outcome.

Some scientists are alarmed by this prospect. It would lead to little change in research policy itself: Schröder would be unlikely to hand over the research ministry to the Greens, as they wished. However, the junior coalition partner would probably be given the environment ministry: bad news for the biotechnology and nuclear industries.

As the Greens have edged closer to power they have shed some of their more uncompromising positions. Their election platform still calls for a complete ban on field trials of genetically modified crops, however, as well as a tightening of gene-related laws. When Joschka

Fischer, now leader of the Green party, became environment minister in Hessen in 1985, he managed to delay for years permission for Hoechst to open a plant making genetically engineered human insulin. He also closed Germany's only plutonium reprocessing plant.

And recently, Rainer Steenblock, Green environment minister in SPD-Green Schleswig-Holstein, opposed planned field trials of genetically modified crops run by AgrEvo (*Nature* **394**, 819; 1998).

As senior coalition partner, the SDP would oppose any move by a Green environment minister to outlaw such trials — but it would have to balance the strength of such opposition against any threat to the stability of the coalition government. A Green environment minister could also expedite the closure of nuclear power stations, with serious implications for German energy policy.

Science and technology issues may have got lost in the election campaign, but they will return to the fore if an SPD-Green government is the outcome. □



Panel urges caution on genetic testing for mental disorders

[LONDON] Britain's main bioethics advisory panel has issued a strong warning against attempts to use genetic screening to predict an individual's susceptibility to common mental health disorders such as schizophrenia and Alzheimer's disease.

Aware that companies may soon be tempted to offer 'over the counter' tests for such diseases, based on the identification of susceptibility genes, the panel suggests that regulations could be needed to replace the voluntary constraints now in force.

But it also says that, in view of the potential importance of genetic studies in treating such diseases, there should be no bar on using people who are mentally incapacitated as research subjects, provided that this is carried out with "the appropriate safeguards".

The conclusions come in a report published this week by a working party of the Nuffield Council on Bioethics, the body set up in 1991 as Britain's national forum for identifying ethical questions raised by advances in biological and medical research.

The report, *Mental disorders and genetics: the ethical context*, says there has been great progress in recent years in refining genetic tests for those relatively rare diseases, such as Huntington's disease and fragile X syndrome, which are generally known to be due to a defect in a single gene. In such cases, it says, tests have a high predictive value clearly relevant to decisions about treatment.

It also points out that researchers working on more complex diseases have frequently identified genetic mutations associated with these diseases, even if almost every such locus is still the subject of scientific controversy. Indeed, the report says that the difficulty of identifying reproducible gene localizations in common mental diseases "represents a key scientific discovery in its own right".

But the working party, which was chaired by Dame Fiona Caldicott, principal of Somerville College, Oxford, and immediate past president of the Royal College of Psychiatrists, warns against over-hasty attempts to use the success of screening tests for single-gene diseases as a basis for promoting the likely value of tests for the others.

In a letter to *Nature* (see page 317), the members of the working party say that, even if a number of susceptibility genes were identified for a particular disorder, "without an understanding of their interaction, they would not be adequate for predicting individual risk in a clinical setting".

Their report identifies Alzheimer's dis-

ease in particular as a case in which, even though there is good evidence of linkage to a particular allele of the apolipoprotein E gene, the clinical and predictive value of this knowledge remains low because other, unknown factors are likely to be involved.

"For such diseases, people may want to look for [genetic mutations] because it may be useful in helping us understand the disease," says working party member Martin Richards, director of the Centre for Family Studies at the University of Cambridge. "But as a public health measure, we see no future in that type of testing in the near future."

Committee members admit that there was a spectrum of views on the long-term potential of genetic tests for multifactorial diseases. "I would be a little more optimistic than some," says Peter McGuffin, professor of psychological medicine at the University of Wales.

But McGuffin, who will shortly become director of the Medical Research Council's Social, Genetic and Developmental Psychiatry Research Centre, also says that disagreement tended to be on scientific grounds rather than over ethical principles.

The Nuffield report is likely to set the framework for more detailed discussion of technical developments by bodies such as the Advisory Committee on Genetic Testing. It

sets its face firmly against "geneticization", which it describes as "an increasing tendency" on the part of some researchers "to view human beings essentially as gene carriers".

It also says that predictive genetic testing and testing for carrier status for mental disorders in children should "be strongly discouraged", and that testing in adoption cases "raises important and complex issues which require appropriate guidance".

And it argues that the present voluntary system of approval for directly marketed tests "is likely to prove insufficient", suggesting that if monitoring of the marketing of such tests and the use of their results shows up adverse consequences, the UK government should seek to impose national or international regulation of such tests.

The main thrust of the report is likely to be endorsed by many of those concerned with the welfare of individuals suffering from genetic diseases.

"I am very glad someone has drawn attention to the difference between the situation facing those at risk from single-gene diseases and multifactorial genetic diseases," says Alistair Kent, director of the Genetic Interest Group. "The single-gene model has served people well, but is not necessarily the way in which we should approach more complex diseases."

David Dickson

US gene-therapy proposals come under fire

[WASHINGTON] An advocacy group is challenging two proposed *in utero* gene-therapy experiments due to be discussed today (24 September) and tomorrow by the Recombinant DNA Advisory Committee (RAC) of the National Institutes of Health.

In a letter sent to the RAC last week, the Boston-based Council for Responsible Genetics charges that the *in utero* method, proposed by gene-therapy pioneer French Anderson, professor of biochemistry and paediatrics at the University of Southern California School of Medicine, leads "farther down the slippery slope to eugenics".

But Anderson said on Monday that the council's criticism was "welcome", and

that, while he judges the risk of germline modification in his experiments to be minute, he has submitted his proposals "in order to have a public debate".

In the council's letter to the RAC, Wendy McGoodwin, its executive director, warns that the proposed experiments risk making modifications to the fetal germ line that would be passed on to succeeding generations.

The RAC currently refuses to consider proposals aimed at modifying the germ line. The Anderson proposals do not seek to do this, but inadvertent germline modification is an admitted risk of *in utero* therapy. McGoodwin urges the committee "to tell Dr

Anderson that the RAC does not consider *in utero* gene therapy to be an acceptable subject for research".

The letter is among 70 pages of public comments, most of them negative, that the RAC has received in response to a public notice of the meeting.

Of Anderson's *in utero* proposals, one aims to treat α -thalassaemia, an error of haemoglobin synthesis, and the other severe combined immunodeficiency caused by adenosine deaminase deficiency (see *Nature* **395**, 8; 1998). Neither protocol is ready for use in humans. At least two years of vector development and animal studies are necessary first, says Anderson. **Meredith Wadman**

France to strengthen regional centres...

[PARIS] The French government has announced plans for a multi-billion-dollar regional development project for higher research and education, known as the 'University of the Third Millennium' (U3M). One of its goals will be the creation of a national network of centres of research excellence with regional universities as the nodes.

The project, to run from 2000 to 2006, is part of a vast multi-annual programme organized by an interministerial committee known as CIADT, which is responsible for planning public infrastructure projects. CIADT's costs are met equally by the state and regional government. Such programmes have traditionally been given a high political priority in France, and are generally adhered to, irrespective of changes of government.

U3M will be organized along the same lines as its predecessor, University 2000 (U2000), a FF42 billion (\$7.4 billion) programme launched by Lionel Jospin, the present prime minister, during his spell as education minister from 1988 to 1992. U2000 has resulted in the construction of about 1.5 million square metres of university space.



Allègre: has scope to reallocate resources.

Claude Allègre, the minister for national education and research, said last week that the costs of U3M would be at least as significant as those of U2000, but that its goals would be very different. The main emphasis would be on research infrastructure, such as high-speed computer networks and new biotechnology centres.

Indeed, research was almost absent from U2000, which was targeted towards refurbishing France's universities, after decades of neglect, and creating space to absorb a massive rise in student numbers. Priority was also given to regional development to offset the historic centralization of higher education in the Paris region.

But demographic changes mean that student numbers are set to plummet, freeing up resources for other goals. These will be drafted by a new national 'strategic committee'

chaired by Guy Aubert, a former director general of Centre National de la Recherche Scientifique (CNRS). Committee members include several university vice-chancellors and scientists such as Jean-Marie Lehn, winner of the 1987 Nobel prize for chemistry.

Proposals will be drawn up in parallel by regional groups of university and local-government officials. A detailed proposal for U3M is expected to be completed by June next year.

"We are no longer in a phase of expansion of student numbers, during which the quantitative needs were prevalent," says Edouard Brezin, chairman of the board of directors of the CNRS and a member of the new committee. "The emphasis on research facilities and quality will be much more important now."

In particular, Allègre sees the promise of a large investment in research infrastructure as a means of building on excellent existing regional research universities — for example, at Strasbourg, Compiègne and Grenoble — to create large, research-based technology universities based on US models such as the Massachusetts Institute of Technology.

The government recently allocated FF1 billion to create a national technology network to bring together public and private laboratories in key sectors (see *Nature* 393, 203; 1998), and industrial development will be a "priority" within U3M, says Aubert. Under preliminary proposals, a national network of biotechnology science parks would be created, for example, centred on the new Gépôle centre at Evry near Paris (see *Nature* 392, 214; 1998).

More broadly, the programme seems to be a reinforcement of the policy of creating regional research 'pôles', which are federations of laboratories that bring together scientists from local universities, research centres, hospitals and industry to create a critical mass in a research project at a single site.

These would be backed up by the construction of shared national science facilities. Allègre has asked research organizations to submit a 'wish list' of medium-sized equipment. Among those he has earmarked as priorities are the creation of a supercomputing network, the building of new structural analysis centres, and an upgrade of the national research Internet network, Renater.

Allègre will benefit from the fact that more than half of France's research and teaching staff are due to retire over the next decade, providing extra scope for reallocating priorities and resources.

"The idea of defining locations for national institutes in various disciplines, which would receive the means needed for modern research, rather than spreading resources thinly, certainly goes in a welcome direction," says Brezin.

Declan Butler

Biology federation urged to widen focus

[WASHINGTON] The US Federation of American Societies for Experimental Biology (FASEB) is being pressed by some of its 17 affiliated societies to work harder to represent the views of individual investigators on professional issues, as well as lobbying for research funds.

At a ten-yearly retreat to assess the strategy of the federation, held two weeks ago at Potomac, Maryland, representatives of the societies said they would like FASEB to pursue issues such as graduate training and the quality of peer review with the same energy with which it argues for more funding for biomedical research.

At a similar retreat in 1989, FASEB was effectively reconstituted as the public affairs arm of its member societies. It has since expanded from six societies with 26,000 members to 17 societies with 56,000 members, and has emerged as the main public voice of US biomedical researchers.

David Kaufman, professor of pathology and laboratory medicine at the University of North Carolina and the new vice-president of FASEB, will take over as its president next July. He says he would like FASEB to develop policies on such issues as the burden of regulation on scientists, restrictions that affect their use of patented research tools and data, scientific integrity, and career development for young scientists.

"It is important to me that we be seen not only to be raising funds for research, but

also in an advocacy role for the individual scientist," Kaufman says. He acknowledges that FASEB had "done a great deal" already in these areas, but often on a reactive basis. "My goal is not just to do it reactively, but to think about these things and develop policies on them at an early stage, perhaps bringing issues to the fore."

FASEB staff point out that it is already active in policy areas. Last year, for example, it produced a consensus report on the need for reform of graduate education for life scientists. But according to one official who attended the Potomac retreat, several society representatives said that FASEB should place more emphasis on professional issues. "They wanted us to focus more broadly — although, to tell you the truth, I thought we were doing that already," says the official.

FASEB has scored successes in recent years by calling for increases in funding at the National Institutes of Health that were then implemented by appropriators in Congress. The societies may now feel that this success can be extended to other areas of professional interest to their members. But Michael Jackson, executive director of FASEB, says it has already sought to address such issues. The reason some people may have the impression that FASEB is fixated with money for research, he says, is that the funding process recurs annually. "With other issues, once they are addressed, they stay addressed," he says.

Colin Macilwain

...and more money for Paris laboratories

[PARIS] There are plans for increased investment in the research infrastructure of Paris, according to preliminary proposals for the multi-billion-franc scheme known as the 'University of the Third Millennium' (U3M). The proposals were discussed at a meeting of the programme's strategic committee last week (see opposite page).

The greater Paris area—Ile-de-France—has been the main loser in recent national schemes to expand research and higher education. Many of its universities and laboratories are in a dire state of decay.

Immediate priorities are the controversial multi-million-franc removal of asbestos from the sprawling Jussieu campus in the Latin quarter (see *Nature* **382**, 291; 1996), and the construction of a university beside the new Library of France in the south-east of Paris. The networking electronic integration of the new library with the university library system is also likely to be given high priority. There are also plans for investment in research facilities at the four new universities created just outside Paris by U3M's predecessor, 'University 2000'.

A renewal of efforts to develop the capital's research activities now also seems likely after a decade of separate government efforts to decentralize research from Ile-de-France. The area, which includes many government-

funded research laboratories, accounts for less than a fifth of France's population and only a quarter of its student population, but carries out more than half the country's publicly funded research.

In 1992, the government demanded that thousands of research posts be transferred from Paris to regional centres by the end of the decade (see *Nature* **356**, 373; 1992). The target set was to reduce the percentage of researchers in Paris to 40 per cent.

The Centre National de la Recherche Scientifique (CNRS) has moved 1,200 posts out of Paris, and, while 51.5 per cent of CNRS researchers worked in Ile-de-France in 1991, now less than 46 per cent do. This represents a fall of 8.7 per cent, whereas overall the number of CNRS researchers has grown by 4.5 per cent over the same period.

CNRS has achieved this shift largely through an affirmative employment policy, recruiting two researchers to regional centres for every one in the capital, and by investing in new regional centres.

Between 1990 and 1998 it spent FF1.2 billion (US\$212 million) on joint regional initiatives with the state and local government, a sum matched by these bodies. It created 31 centres of excellence outside Paris, including one in electronics in Lille, materials research in Bordeaux, structural biology at Grenoble,

and even a centre for taste research at Dijon, the home of French mustard.

But CNRS is now reconsidering its strategy. An internal working group on regional affairs recently submitted a report to Catherine Bréchnignac, the director general of CNRS, asking whether it is time to reconsider its policy of differential recruitment.

The working group, chaired by Jacques Sevin, director of strategy at CNRS, argues that the ageing of the Paris research population provides a case for renewed recruitment and that, unless it receives renewed attention, its science base is likely to suffer. Ile-de-France is top of a new ranking of research activities (see below).

The group argues that, rather than decentralization from Ile-de-France, the development of regional metropolises should be the basis of the development strategy of CNRS.

The rehabilitation of Paris is likely to be included in the CNRS's regional policy, to be made public by Bréchnignac later this year.

Sevin adds that regional development remains one of the biggest policy issues in CNRS. "The big work has been done, the task now is to consolidate," he says. He adds that CNRS needs to organize a critical mass of scientists in new regional centres and to devote more investment to equipment than to constructing new laboratories. **DeclanButler**

Map of research activities highlights disparity across European regions

[PARIS] The geographical distribution and abundance of research is a burning political issue in France (see page opposite and above), and discussion of it at the European level seems likely to be stimulated by a map of research activity across Europe published by the Paris-based Observatoire des Sciences et des Techniques (OST).

More than 80 per cent of the European Union's scientific output in terms of papers and patents is accounted for by just 67 regions on the map. The Paris area came top of the league, followed by London, Munich and Düsseldorf. Paris and London alone accounted for almost 10 per cent of Europe's scientific output.

The map (right) shows clearly that research activity is highest in the economically strong 'banana' extending along the axis of Milan, Rotterdam and London, with a skew towards

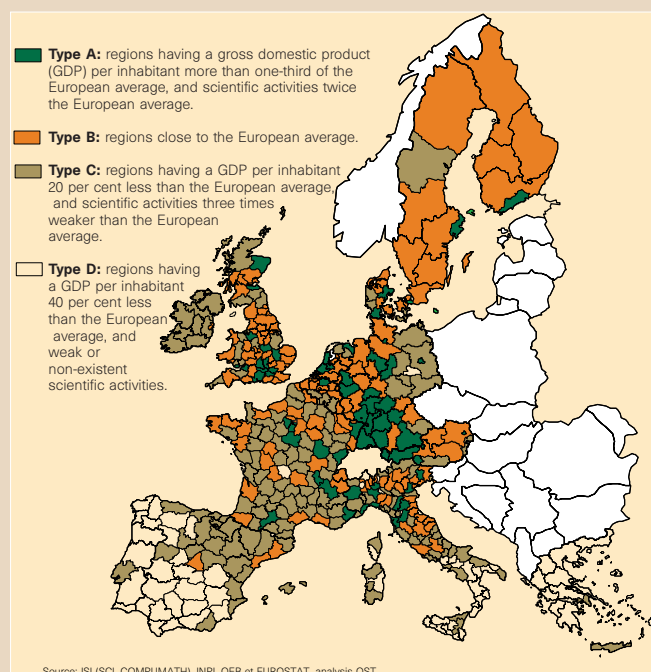
Paris and Lyons. Other striking features are that Italy and Germany are both split into two halves, one half rich in research, the other poor.

Research activity in Scotland is also revealed to be much stronger than might be widely believed, adds Remi Barré, director of OST, adding that Nordic countries also have strong research according to the map.

"The aim is to better understand the regional reality of scientific activity in Europe," says Barré, arguing that increasingly the regions, with their own substantial funds, are major players in research.

Barré cautions that the map is a simplified version of a detailed OST analysis, co-financed by the European Commission, of the research activities of the 445 regions in the member states of the European Union.

Regions were selected using



the commission's NUTS scale, which classes regions from NUTS 0 (a country) to NUTS 5 (a village). The OST analysis mostly included regions the size of NUTS

2 or 3 (roughly the size of a county, depending on the country), although tiny Luxembourg was counted as a single region, NUTS 0. **D.B.**

India blocks panel hearings on impact of dam construction

[NEW DELHI] The Indian government has stirred up controversy by withdrawing permission at the last minute for the Cape Town-based independent World Commission on Dams (WCD) to hold a set of public hearings in India on large dams. The meeting will now be held in Cape Town.

The WCD was formed in February at the initiative of the World Bank and the World Conservation Union. It aims to start a dialogue on developing dams that will have minimal environmental impact. Its worldwide membership includes governments, non-government environmentalist groups and indigenous people's organizations.

The Indian hearings were to have been the first in a series around the world aimed at developing a set of guidelines for the planning, design and construction of environmentally safe dams.

Having obtained government clearance, the commission had scheduled Indian hearings in Bhopal and New Delhi this week (19–23 September). But on 11 September, WCD chairman Kader Asmal, South Africa's minister for water resources and forestry, was told that it was not an "opportune" time to visit India to conduct the hearings.

Indian officials say the government asked the commission to "postpone" its visit, as it did not want the WCD hearings to interfere with a legal case over the Sardar Sarovar dam project which is due before the Indian supreme court later this month.

The massive dam across the River Narmada has been embroiled in bitter controversy for several years between the Gujarat government, which is building the dam, and environmental and human-rights activists. These are led by sociologist and WCD member Medha Patkar of Narmada Bachao Andolan (Save Narmada), which claims that the Gujarat government has resettled only 4,500 of the 120,000 people displaced by the dam.

A promise by the commission that the hearings would exclude reference to the Narmada dam project has not persuaded New Delhi to alter its stand.

Patkar has alleged that the Indian government cancelled the WCD's visit under pressure from Gujarat politicians who believe the commission consists mainly of anti-dam activists. Two weeks ago, the Gujarat state legislature passed a resolution describing the WCD's proposed visit as "an intrusion into our internal affairs".

But according to Asmal, the commission seeks to bring "a more responsible approach to investments in large dam projects" by conducting a comprehensive review of their costs and benefits.

K. S. Jayaraman

Edgy Japan wants a new reconnaissance satellite

[TOKYO] Japan is looking into launching a reconnaissance satellite to improve its ability to gather security information. The move comes despite a long-standing parliamentary resolution criticizing the 'militarization' of space, and follows last month's firing of a North Korean missile over Japanese territory into the Pacific Ocean.

Fears that North Korea had test-launched a three-stage ballistic missile on 31 August put Japan on a high-security alert. North Korea says it was actually launching a 'singing satellite' for peaceful purposes — such as broadcasting patriotic songs in praise of 'dear leader' Kim Jong-il. It is believed to have used a multi-stage missile with an estimated range of 2,000 kilometres.

The US State Department last week concluded that the suspected third stage of a missile actually was a small satellite, but that it failed to achieve orbit. Even if it was, Japan still views North Korea's action — firing a long-range missile over Japan without warning — as a threat to its security.

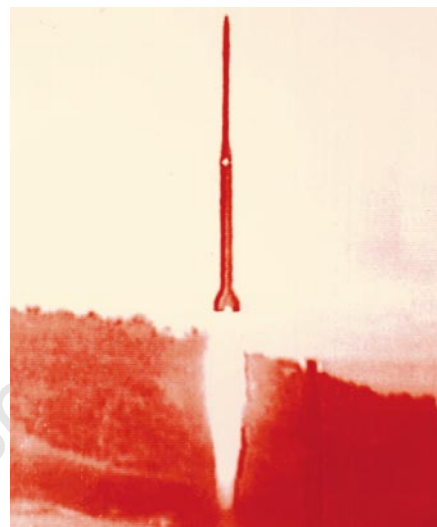
The Japanese government responded by freezing various aid packages, including 135 billion yen (US\$1 billion) for an international project to build light-water reactors in North Korea. The project, carried out under the auspices of the Korean Peninsular Energy Development Organization, started in 1994 and was meant to stop North Korea pursuing its suspected nuclear weapons programme.

The rocket launch led Japan's ruling Liberal Democratic Party (LDP) to set up a working group to discuss the introduction of a multi-purpose satellite system with reconnaissance capabilities.

"We have instructed the ministries and agencies concerned to study types of satellite systems suitable for carrying out reconnaissance activities," said Prime Minister Keizo Obuchi at a recent LDP meeting in Tokyo.

Particular attention has been given to the National Space Development Agency's Advanced Land Observation Satellite (ALOS), scheduled for launch in 2003 to carry out a mapping, regional observation, disaster observation and earth resources survey. Ironically, ALOS, which was originally called the High-Resolution Land Observation Satellite, changed its name after the agency thought it sounded like a reconnaissance satellite with military purposes.

"Although we are obtaining information from all space-related sources under the respective ministries and agencies, the technology obtained through the development of ALOS would certainly be a great advantage for planning a possible model for Japan's reconnaissance satellite," says Hideo Fun-



Hit or missile: was the North Korean rocket that buzzed Japan just trying to spread some music?

abashi, who is responsible for space research at the Science and Technology Agency (STA).

The United States is willing to cooperate in the development of such satellites. "It is still not known whether the cooperation would be at the technical level, but there will certainly be an exchange of information," says Funabashi.

The main sticking point could be a resolution passed by the Diet (Japan's parliament) in 1969 on the non-militarization of space. But unusually strong support from opposition parties, including the LDP's main rival the Democratic Party of Japan, may ease some of its restrictions.

As a result of a bilateral security meeting in New York last Sunday (20 September), Japan and the United States have agreed to begin joint research on the TMD system — a comprehensive defence system for intercepting ballistic missiles — which they have been discussing since 1993 when North Korea test-fired a missile into the Sea of Japan.

Japan is expected to begin the collaboration in the fiscal year 1999, which starts in April. Parliamentary approval for the project, and the allocation of an additional budget estimated at ¥5–¥10 billion, is expected to be made official by the end of this year. But the main cost of the project is expected to be between ¥20 billion and ¥30 billion for the five-year period, which is causing concern given Japan's poor economic prospects.

There is also anxiety about instability in the Defence Agency, which has been rocked by a financial scandal. Agency officials added insult to injury by trying to dispose of documents related to the scandal amid the chaos following the missile launch. **Asako Saegusa**

SOHO returns, Mars Surveyor hiccups

[WASHINGTON] Ground controllers regained command of the errant Solar and Heliospheric Observatory (SOHO) last week for the first time in almost three months. Their achievement has raised hopes that scientific use of the European-US satellite could resume within the next two months.

After weeks of slowly thawing fuel lines connected to SOHO's onboard thrusters, a team of European and US engineers working at NASA's Goddard Space Flight Center in Maryland successfully commanded the spacecraft to point its onboard solar power arrays toward the sun on 16–20 September.

Next will begin check-outs of key engineering subsystems and SOHO's 12 scientific instruments, which suffered extreme heat and cold during the period in which the spacecraft was uncontrolled. Mario Acuña, project scientist for the International Solar-Terrestrial Physics programme at Goddard, says that temperatures on some instruments went beyond the range that onboard sensors could record.

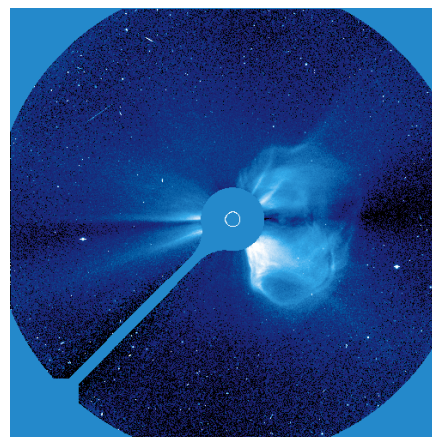
Still, he says, "the mood is one of optimism" that SOHO will resume scientific observations. If all the instruments work

properly, that could happen in as little as four to six weeks, according to project engineers.

An accident investigation board determined earlier this month that the loss of SOHO in June was due to mistakes made by ground controllers, including an erroneous decision that disabled part of the spacecraft's autonomous failure detection system. Ironically, the day after SOHO was recovered, the Mars Global Surveyor (MGS) currently in orbit around Mars experienced its own brief shutdown caused by a faulty computer instruction sent up from the ground.

Engineers at NASA's Jet Propulsion Laboratory had been preparing for a propulsive manoeuvre to begin lowering the Surveyor's orbit around the planet when a bad software command turned the craft's solar power panels at the wrong angle to the sun. That made onboard batteries start draining power. In this case, however, the spacecraft's automatic fault protection system caught the mistake and aborted the manoeuvre when power fell below 50 per cent.

The fact that a faulty software command survived several levels of review on the ground before being transmitted to the



Former glory: SOHO's view of a coronal mass ejection from the Sun last April.

spacecraft is "a little disconcerting", admits MGS project manager Glenn Cunningham. But — unlike SOHO — the MGS's fault protection system saved the day. "It did exactly what it was supposed to do," says Cunningham. As a result, the Mars spacecraft is safe and stable, and is due to resume lowering its orbit as planned this week. **Tony Reichhardt**

Call for scientific experts to keep State Department up to speed

[WASHINGTON] A dozen well placed technical staff and a \$500,000-a-year external advisory board could revive the scientific and technical capabilities of the US Department of State. So says a preliminary report prepared by the National Academy of Sciences and delivered earlier this month to Madeleine Albright, the secretary of state.

Important issues on science and technology "are not receiving adequate attention within the department", says the report. It was written over the summer by a panel chaired by Robert Frosch of Harvard University's John F. Kennedy School of Government in response to mounting criticism of how the state department handles science-related issues.

Briefing a meeting of the President's Council of Advisors on Science and Technology on the study last week, Frosch said there were plenty of sources of sound advice on science and technology available in Washington, but that the state department was ill-equipped to access it.

"The question is how the state department can develop better internal receptors" for scientific input, Frosch said. "We suggest that the secretary of state asks one of the under-secretaries to take special responsibility for these issues."

Scientists in the United States have long complained that the state department, which is responsible for US foreign affairs, is

ill-equipped to handle scientific and technical questions. But the complaints have sharpened markedly over the past year (see *Nature* 392, 427; 1998).

Critics charge that, under the Clinton administration, the state department's Office of Oceans and International Environmental and Scientific Affairs (OES) has concentrated on the environment at the expense of science.

They also claim that the department's meagre scientific and technical capability is overwhelmed by the range of foreign-policy issues — from international copyright laws to trade disputes involving satellites or genetically modified organisms — which now revolve around science or technology.

The Frosch study says that the state department should hire two or three technical staff in the office of the under-secretary who would be assigned special responsibility for science and technology. It should also develop "several additional clusters" of scientific competence at five of its other Washington bureaux.

The report adds that the department should set up procedures enabling it to second people from science agencies, such as the National Science Foundation, to fill science counsellor positions at US embassies with suitably qualified staff. It suggests an array of possible mechanisms to garner advice on science and technology.

The most comprehensive of these would be a formal advisory committee modelled on the Defense Science Board, which advises the Pentagon.

The panel was hesitant in recommending the appointment of a science advisor to the secretary. Glenn Schweitzer, the study director, notes that interest groups such as women and labour unions have pressed for similar appointments to represent their particular interests. "We're reluctant to line up with them just yet," Schweitzer says.

The department is recruiting a science adviser for Melinda Kimble, the assistant secretary for OES (see *Nature* 393, 612; 1998). But Frosch says this will have little effect on the department as a whole.

The interim Frosch report was produced quickly to help the state department prepare its budget for 2000. In a letter to Albright, Bruce Alberts, the president of the National Academy of Sciences, said the report's recommendations could be implemented with the establishment of about 12 positions and, at most, \$500,000 for the advisory structure. "These seem modest investments given the stakes involved," Alberts wrote.

Alberts also asked for a meeting with Albright to help direct a fuller report from the Frosch panel, due in a year's time. Although she requested the study, the most senior official to meet the Frosch panel was her deputy, Strobe Talbot. **Colin Macilwain**

Brussels rewrites mission of 'Eurolabs'

[MUNICH] Europe's much criticized Joint Research Centre (JRC) was relaunched this week. Its new mission is explicitly to support European Union (EU) policies and meet the needs of European citizens.

The JRC brings together seven institutes, including 15 centres and reference laboratories, and acts as an EU reference centre for science and technology. Its largest site, at Ispra in northern Italy, hosts the Institute for Advanced Materials; the Institute for Systems, Informatics and Safety; the Space Applications Institute; and the Environment Institute. The Institute for Reference Materials and Measurements is in Geel, Belgium; the Institute for Transuranium Elements is in Germany; and the Institute for Prospective Technological Studies in Seville, Spain. The institutes employ 1,820 staff, 80 per cent of whom are scientists or engineers.

The new mission statement, which reflects the European Commission's (EC) more socially orientated fifth framework programme of research (FP5), due to start next year, is a response to criticisms that the JRC had become inefficient, expensive and of uncertain scientific merit.

It is accompanied by a planned rationalization of its over-diversified scientific and technological programmes, and follows a recent decision to hire more temporary staff in order to increase its flexibility.

The JRC — founded in the late 1950s as a centre for the development of nuclear power — suffered when nuclear power fell from political favour in the 1980s. In the absence of a clear post-nuclear mission, the number of add-on units grew in an undirected way.

During the past few years the JRC has

worked hard to bring its house in order. This follows a decision made in 1993 by the Council of Ministers, following a drive led by Britain and Germany, that the JRC should obtain a significant part of its budget through competition. In fact the JRC has more or less reached the targets set in the fourth framework programme (1994–1998) — 22 per cent competitive money for its non-nuclear programme and 15 per cent for its nuclear programme.

In future the centres will have only 40 per cent permanent staff, with 35 per cent on renewable five-year contracts and 25 per cent on three-year non-renewable contracts. Many staff hired in the 1960s are starting to retire, making this target feasible in the next ten years. Over 200 new staff were hired last year, all but seven on temporary contracts.

Herbert Allgeier, who directs the JRC's advanced technology programme, was brought in as director general in February to implement the organizational changes. "The JRC must be reconfirmed as an integral part of EU policy-making — and not apologetically defended," he says.

Allgeier intends to start decommissioning research reactors which have not operated for many years. He also plans to bring some of the Environment Institute's units (relating to food and drink quality, control of chemicals and development of alternatives to animal research) into a new Institute for Health and Consumer Protection. "Institutes were involved in too many different tasks," says Allgeier. "Their tasks will now be rationalized to fit the new mission."

New tasks, such as the validation of methods developed to detect genetically modified

organisms (GMOs) in foods, will be added to the list of tasks. Elke Anklam, who heads the JRC's unit for food safety and pioneered the new GMO studies, believes that the reorganization will help to make the JRC serve the EU's aims more efficiently.

The commission is keen that the JRC should maintain its budget for 1999–2002. This is currently under negotiation between the European Parliament, which shares the EC's view, and the Council of Ministers, representing the member states, which wants to reduce it by at least 15 per cent.

"If the council position wins through we will have to close down an entire institute," says Allgeier. All technical support to the European fusion programme will have to be stopped, he says, in addition to other important programmes.

The council and the parliament are also discussing the level of funds the JRC should be asked to raise from outside sources, a matter of considerable controversy. Parliament, as well as the commission and some member states, sees a contradiction between the role of the JRC as an institution at the service of the commission and the demand that it should find money to pay for this service from elsewhere. In order to meet targets in FP4, some JRC institutes had felt under pressure to become involved in projects that would fit criteria for external grants even though they were not necessarily central to their mission, says Allgeier.

He welcomes the continued demand for the JRC to win some of its budget through competitive grants because of the advantages it brings through networking with national research institutes. "But it is nonsense to say we must do it to increase our budget," he says.

This is also the position of Parliament. But Parliament disagrees with Allgeier's proposal, discussed by the JRC's board of governors in Seville this week, to restrict the role of the board to strategic and scientific advice, and monitoring of the institutes.

Allgeier is unhappy with the board's involvement in what he describes as "micro-management". But Eryl McNally, a socialist member of the European Parliament from the UK who is on the Parliament's research committee, says that the board should be involved in detailed management because it needs to be able to follow procedures actively to make recommendations about funding.

The board will also discuss a proposed change in methods of evaluating institutes. The commission would like to be controlled by the individual directorates in Brussels which ask for particular bits of work. A spokesman says that evaluation should not focus on counting publications and patents, but on the efficiency with which contracted work is carried out.

Alison Abbott

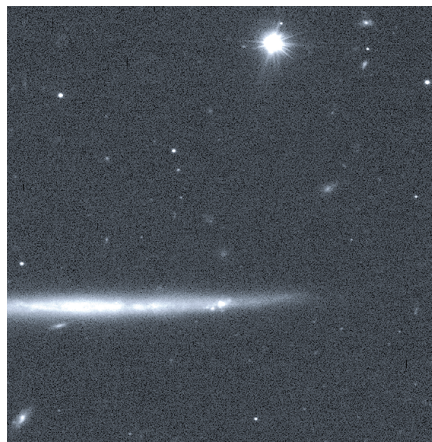
New Chile observatory shows its mettle

[MUNICH] The European Southern Observatory (ESO) has released new pictures of ESO342 – G017, a spiral galaxy outside the Milky Way in the southern constellation of Sagittarius, 400 million light years from Earth.

They are among the first observations from ESO's 8.2-metre Very Large Telescope on Mount Paranal in Chile. More distant galaxies can be seen in the background (right). The brighter parts of the flat disk are believed to be star-forming regions.

The release of the pictures, taken last month, coincided with a meeting of the observatory's governing council last week that formally approved ESO's involvement in a new international large millimetre astronomy observatory in the southern hemisphere (see *Nature* 388, 412; 1997).

The new Large Southern Array/Millimetre Array will merge European and



US plans for a large millimetre array and should cut costs to about US\$350 million. It will provide a millimetre counterpart to the Very Large Telescope and the NASA/ESA Hubble Space Telescope.

A. A.

UK life science students seek better deal

[LONDON] Britain's life scientists are preparing to lobby the government to double the value of a doctoral student's stipend and add an optional foundation year to existing three-year PhD programmes.

The moves follow concerns expressed by universities, research councils and industry that the quality of applicants for PhD courses in the life sciences is falling, and that urgent measures are needed to attract able students to research careers.

Life scientists also want to set up a body to lobby the government on their behalf. This stems partly from a belief that existing organizations have failed to communicate their concerns to government effectively.

The decisions came at the end of a day-long conference in London earlier this month on the training of life scientists. It was organized by the UK Life Sciences Committee, a year-old umbrella group made up of most of Britain's learned societies in the life sciences.

The conference covered most aspects of life-science research training. But a key issue was the urgent need to raise the level of a government-funded life science student's PhD stipend, bringing it in line with that offered by the Wellcome Trust.

Despite being topped up by £1,000 (US\$1,680) after the government's recent Comprehensive Spending Review, the £6,545 annual PhD stipend is still considerably short of the average starting salary for graduate jobs in the United Kingdom.

Delegate after delegate rose to say that many bright students, already burdened with undergraduate debts, were unwilling to devote a further three — possibly four — years earning "slave labour wages". A consensus emerged that the stipend needed to be doubled, even if this meant a cut in the numbers of PhD studentships.

John Lackie of the Yamanouchi Research Institute in Oxford said: "We produce far too many research students, and far too many of low calibre. A bigger stipend is needed. Industry needs excellence, not mediocrity."

This call was echoed by Malcolm Skingle of Glaxo Wellcome, who said his company had recently hired recruitment consultants to "scour Europe", as high-quality postdocs were becoming difficult to find in Britain.

Peter Swann, professor of biochemistry and molecular biology at University College London, said that he found himself offering PhD studentships to graduates whom he had earlier rejected. This was to fill vacancies created by better-qualified students who had decided to pull out.

Bob Price, director of Human and Corporate Resources at the Biotechnology and Biological Sciences Research Council (BBSRC), said that the BBSRC wanted to raise PhD stipends in the life sciences, but was ham-

strung by government policy that stipends remain equal across all research councils.

BBSRC's governing council would shortly be discussing the issue, said Price, adding that stipends should be set according to market demand. "The market for a life scientist is not the same as that for a physical scientist. If you are a space scientist, your options are not wide. But it is different in life science."

Delegates overwhelmingly endorsed the need for a foundation year as another way of attracting — and keeping — able students. There was considerable support for a four-year PhD with a "get-out" clause in which students could leave after the first year with a master's degree. This would be similar to the existing research master's degree, MRes.

Some delegates were concerned that the extra year would leave British PhDs older, and thus potentially less marketable for non-academic careers. But others pointed out that at 27, a postdoc from Scotland who completed the extra foundation year was still much younger than a US or European counterpart.

The foundation year would be designed to train students in research skills, as well as expose them to a range of research projects under different supervisors so that their final

choice could be an informed one.

In addition, the foundation year would give a student "transferable skills" relevant to other employers — such as information technology and public speaking.

The conference debated at length whether more time should be

given to transferable skills. Postgraduate departments offer such skills partly because of student demand, and because they recognize that many students will end up in industry, rather than academia.



Martin Raff: original thinking is the key.



The overwhelming view of the conference was that the research content of a PhD should not be diluted. Most delegates acknowledged the value of transferable skills, but warned of the dangers of overemphasizing them.

"If we're not careful, we'll have lots of students with lots of communication skills, but with very little to communicate," said Mike Carter, professor of biological sciences at the University of Surrey.

By contrast, some delegates complained that British postgraduate education was too focused, and that many PhD students completed their three years without demonstrating evidence of what one delegate described as "the intellectual experience", or, indeed, of having read beyond their specialist subject.

"There's a lot of emphasis on training for industry," said Martin Raff, chairman of the UK Life Sciences Committee. "What is key is learning to think originally. Yet few students seem to show evidence of this."

One reason, in Raff's view, was the "anti-intellectual" nature of some current life science research, particularly molecular biology. Raff said he sympathized with students who spent three years decoding the function of a single gene. However, he felt that the era of formula science would pass once the Human Genome Project was complete. **Ehsan Masood**

France to set up co-ordinating committee

[PARIS] The French government has announced the creation of a national coordinating committee for research in the life sciences, to be attached to the ministry of national education, research and technology.

It will be chaired by Nicole Le Dourain, professor at the Collège de France in Paris, and will have twenty other members. These will include representatives of the nine public research organizations with interests in life sciences, among them Claude Griscelli, director-general of the biomedical agency INSERM, and Maxime Schwartz,

director-general of the Institut Pasteur. Other members will be leading scientists such as the geneticist Jean-Louis Mandel.

The committee's task will be to propose and evaluate programmes coordinated among the research agencies, to advise the ministry on strategy, and to set up debates on research issues such as the organization of genome and 'post-genome' research, telemedicine and genetic engineering. The committee will be consultative, but, as well as replying to requests from the ministry, it will issue opinions on subjects of its choosing.

Calls to standardize CO₂ emissions gain further support

[LONDON] The European Parliament and states belonging to the Non-Aligned Movement (NAM) have added their voices to calls for the world's carbon dioxide emissions to converge at a single per capita level by the next century.

The parliament passed a resolution to this effect last week. A summit of NAM member states held in Durban, South Africa, at the beginning of the month also called for individual emissions entitlements to be distributed on a per capita basis.

The principle of emissions trading will be high on the agenda of the conference of parties to the United Nations climate convention in Buenos Aires in November.

Under the Kyoto climate protocol, signed last year, developed countries have to reduce greenhouse gas emissions by 5 per cent from 1990 levels before 2012. Countries that emit above the baseline will be able to trade entitlements from those, such as Russia, that emit below the baseline.

Missouri wins NSF grant for maize genome work

[WASHINGTON] The University of Missouri at Columbia has won an \$11 million grant to establish a maize genome centre under the National Science Foundation's Plant Genome Initiative — a \$40 million-a-year research project which the NSF was directed to perform by Senator Christopher Bond (Republican, Missouri).

Bond chairs the Senate appropriations subcommittee that funds the NSF. The award, the largest grant ever won by the university, arrives little more than a year after the Missouri senator proposed that the agency should start the Plant Genome Initiative to bolster genomic research in crops (see *Nature* 388, 312; 1997).

The grant is one of ten to be awarded in the first year of the initiative; NSF officials say they are not ready to announce the other nine. The university, the NSF and Bond all say that the grant was awarded purely on its scientific merit. The maize genome centre will be housed in a new life sciences building already planned for the university campus. It will work to enhance a maize genome database currently maintained by the US Department of Agriculture.

Plant genome centre set up in France

[PARIS] French public research agencies and agrofood companies last week agreed to pool their efforts in plant genome research, in particular through the creation of a joint

genomics centre at the Génomopole biotechnology park at Evry near Paris. Under the terms of the preliminary agreement, a new joint research programme would be funded equally by industry, the state and the French National Institute of Agronomic Research (INRA).

Other research organizations will also be involved, and participating companies include Rhône-Poulenc and Biogemma, the plant biotechnology subsidiary of Limagrain and Pau Euralis. The costs of the programme — known as Génomplante — have not been revealed, but will require a "high level of investment", according to Paul Vialle, director general of INRA.

Developing countries told to cut ozone chemicals

[LONDON] The responsibility to restore ozone levels rests largely with developing countries and economies in transition, according to Klaus Töpfer, executive director of the United Nations Environment Programme (UNEP).

The 1987 Montreal protocol on ozone-depleting substances allowed developing countries ten years before they needed to phase out such chemicals. "This period of grace is over," said Töpfer, speaking on World Ozone Day on 16 September. He added that phasing out had to begin by 1 July 1999.

Ten years ago, developing countries and the Russian federation accounted for just 15 per cent of the worldwide production of ozone-depleting chlorofluorocarbons. In 1996, that figure reached 80 per cent, according to a UNEP statement. The consumption of chlorofluorocarbons in industrialized countries during the same period has dropped from one million tonnes per year to 15,000 tonnes.

May and Craig named as Balzan prize-winners

[LONDON] Britain's chief scientific adviser, Sir Robert May, and Harmon Craig of the Scripps Institute for Oceanography in San Diego, have been announced as two of this year's winners of the annual SF500,000 (US\$360,000) Balzan Prizes. The annual awards are made by the International Balzan Foundation, which is based in Italy and Switzerland. They recognize leading personalities from the scientific, cultural and humanitarian fields.

May has been cited for his contributions to the understanding of the relationship between biological diversity and the structure and functioning of the resulting ecosystems, and Craig for his work in geochemistry. The third winner was historian Andrezej Walicki. The prizes will be awarded in November by the president of Italy.

India opens its third reprocessing plant

[NEW DELHI] India's third and largest plutonium reprocessing plant, at Kalpakkam in Tamil Nadu state, was officially inaugurated last week by the prime minister, Atal Behari Vajpayee, who promised the scientists "every help from the government to realize the national mission of nuclear technology". The plant will be able to handle 100 tonnes of spent nuclear fuel a year, separating out the plutonium which India says will be used in fast-breeder reactors to generate electricity.

The two other reprocessing plants are at Mumbai and about 150 km north of there at Tarapur. While the Mumbai plant has been extracting weapons-grade plutonium from spent research-reactor fuel since 1964, the 10-year old Tarapur plant has been reprocessing fuel discharged from power reactors. Two of the five devices tested by India in May this year are believed to have used plutonium that was not classified as weapons grade.

Siberian centres face closure over debts

[MOSCOW] The Siberian branches of the Russian Academy of Sciences are on the verge of being closed down because their joint debt for local utilities is now over 170 million rubles (about \$12 million). Another factor leading to the crisis is that Siberian scientists have received only 44 per cent of the annual budget that was approved by the federal cabinet in Moscow.

In an attempt to save its institutes and laboratories from closure, the Krasnoyarsk scientific centre has decided to give its scientists only 80 per cent of their already decreased salaries. The remaining money will be used to pay for electricity, heating and other services.

Funding alternatives 'won't cut medical costs'

[MUNICH] Reimbursing the costs of complementary or alternative medicine, which is widely accepted in Switzerland, would not result in savings for national health insurance companies, according to a study recently released by the Centre for Economics (WWZ) of the University of Basel.

The study compared adults who take out extra insurance for complementary therapies with those who have only health insurance for conventional treatments. It found that patients mostly used complementary medicine in combination with conventional medicine, rather than relying on it alone. If national health insurance companies were to offer to reimburse their costs generally, says Jürg Sommer of the WWZ, national health costs would rise.

Restrict genetic susceptibility tests

Sir— Given the extent of current research into both genetics and mental disorders, the Nuffield Council on Bioethics has undertaken an inquiry into issues raised when these fields come together. The report, 'Mental disorders and genetics: the ethical context', published on 23 September, is primarily concerned with whole persons and not simply with their genes. This broad and humanistic perspective may be contrasted with a reductionist approach which risks undermining both moral responsibility and social solidarity. Genetic information about mental disorders may alter the way those affected are viewed by others and in particular, the stigma they suffer.

The discovery of mutations associated with the 'traditional' Mendelian single gene disorders has had profound implications for clinical diagnosis and predictive testing. However, the Nuffield inquiry concluded

that genetic tests for the diagnosis of the common mental disorders with more complex causes will not be particularly useful in the near future. Even if a number of susceptibility genes were identified for a particular disorder, the Nuffield Council takes the view that, without an understanding of their interaction, they would not be adequate for predicting individual risk in a clinical setting. It has therefore recommended that genetic testing for susceptibility genes which offer relatively low predictive or diagnostic certainty be discouraged unless and until there is clear medical benefit to the patient.

The Nuffield Council would construe as unethical the exclusion of people with a mental disorder from genetic research, as the identification of associated genes may lead to more effective drug treatments. The view is taken therefore that non-therapeutic research should be considered ethically

acceptable even when it involves people lacking the capacity to consent to participation, subject to strict safeguards. While the best safeguard against new eugenic pressures is freely given, properly informed consent, guidelines for the establishment and maintenance of genetic registers are needed.

Sandy Thomas, (Director), Fiona Caldicott*, (Chairman of the Nuffield Council on Bioethics Working Party on Mental Disorders and Genetics), Chris Barchard*, Rachel Bartlett, John Haldane*, Sonia Hornby*, Peter McGuffin*, Nigel Pleming*, Martin Richards*, Pamela Taylor*, Andrew Wilkie* & Sally Young*

*Nuffield Council on Bioethics,
28 Bedford Square,
London WC1B 3EG, UK*

**Member of the Working Party*

1. *Mental disorders and genetics: the ethical context* (Nuffield Council on Bioethics, London, 1998).

Star chambers will result in injustice

Sir— Jon Turney's generally favourable review of my book *The Baltimore Case* offers the judgement that the affair was "not of enormous significance, except for the victims" (*Nature* 395, 30–31; 1998). That is to miss the point of historical importance of the case, and a major point of the book. The case had great significance for the civil rights of scientists charged with fraud.

Thereza Imanishi-Kari was investigated relentlessly by a congressional subcommittee and by an entity in the US Department of Health and Human Services now named the Office of Research Integrity (ORI). In both venues, she was denied elementary protections of due process, especially the right to see and evaluate evidence and the right to confront and cross-examine witnesses against her. So she could not defend herself effectively. The miscarriage of justice was put right because Imanishi-Kari's tenacious refusal to concede ultimately gained her access to an appeals board that afforded her full due-process rights.

I think it perfectly legitimate — indeed, obligatory — for government authorities to call to account publicly supported scientists who cheat. My quarrel with the investigation of Imanishi-Kari is not primarily, as Turney says, that the authorities went after "the wrong people". It is that the ORI went after Imanishi-Kari in the manner of a Star Chamber. The denial of due-process rights to anyone, innocent

or guilty, is indefensible. The ORI investigators defended their procedures on the grounds that they were not conducting a legal inquiry but engaging in a dialogue between scientists. But, if a criminal penalty was not necessarily at stake, careers and reputations were.

The case has already had a major impact on how such charges in biomedical research are handled in the United States — by, for example, establishing the right of all accused scientists to use the appeals process. The case is deeply instructive for scientists and policymakers in the United States, where the procedures for handling fraud cases are still evolving, and in other countries, where questions of how charges of scientific fraud should be adjudicated are commanding increasing attention. The ORI's original procedures threatened the civil rights of every scientist who might be charged with fraud.

Daniel J. Kevles

*California Institute of Technology (228-77),
Pasadena, California 91125, USA
e-mail: kevles_d@caltech.edu*

Was Moses the first ecologist?

Sir— A recent UK television programme, *Equinox: The Ten Plagues of Egypt* (Channel 4), gave a modern reappraisal of the sequence of biological and meteorological catastrophes recorded in the biblical book of "Exodus".

It was proposed that, within the unavoidable constraints of the language used, the account can be taken as factual — and in particular that the sequence of these events is significant. The trigger was a red algal bloom that killed much of the life in the Nile, especially the fish which normally kept the toad population under control. It was argued that this led, by plausible steps, to plagues of flies and midges and hence to animal and human plagues for which they were vectors. The final catastrophe arose from the storage of wet grain contaminated by locust droppings which led to the poisoning of human and animal food supplies with microtoxins.

The corollary, which the programme did not stress, is the predictability of much of this sequence. Algal blooms must have occurred before (although perhaps not on such an epic scale) and farmers must have been forced to store wet grain in earlier harvests because of bad weather conditions. Observers must have seen and recorded the results. The educated elite must have known what to expect (and Moses had been adopted by Pharaoh's daughter and educated as a prince in the royal court). Armed with this knowledge, he would have been able to predict publicly, to good effect, the exact form each of the biological catastrophes would take, before it occurred.

The mistake was to educate Moses. One can picture the Pharaoh concurring with Oscar Wilde that no good deed goes unpunished.

John Lydon

*School of Biochemistry and Molecular Biology,
University of Leeds, Leeds LS2 9JT, UK*

A home from home for coelacanth

Peter Forey

The discovery of specimens of the coelacanth, *Latimeria chalumnae*, in Indonesian waters raises questions about the geographical distribution and conservation status of this remarkable fish.

On page 335 of this issue¹ Erdmann *et al.* report the existence of living coelacanths off the northeast coast of Sulawesi, Indonesia, some 10,000 km east of their presumed natural home of the Comoros archipelago in the western Indian Ocean (Fig. 1). The circumstances of the initial discovery are astonishing — Erdmann's wife happened to see the instantly recognizable form of a coelacanth being carted through a fish market; Erdmann and colleagues later saw and described a live Indonesian specimen.

The first living coelacanth was caught in 1938 in a shark gill-net off the mouth of the Chalumna River, South Africa, and was named *Latimeria chalumnae* by J. L. B. Smith². The catch raised great excitement, for two reasons. First, this was the only surviving species of a lineage of fishes — the coelacanths — that originated in the Devonian, about 360 million years ago, but was thought to have become extinct in the Upper Cretaceous some 80 million years ago, the date of the youngest fossil recovered. So here was a true 'living fossil'. How had the lineage

survived for that time without leaving any trace in the fossil record?

Second, the theory of relationships prevalent in the late 1930s held that the coelacanth was a direct descendant of Devonian fish-like ancestors of the tetrapods (land-living vertebrates, including ourselves). By studying *Latimeria* it was expected that we might understand something about the physiology, behaviour and ecology of our own remote ancestors during the transition from life in water to life on land.

From modern evolutionary analysis³ it seems that the coelacanths are more distantly related to tetrapods than first thought, and that they are their modern cousins rather than their sisters (lungfishes are now taken to be the closest living relatives of tetrapods). But *Latimeria* may still hold the answers to several questions. It is the only living animal to have a functional intracranial joint: that is, a complete division running through the braincase and separating the nasal organs and eye from the ear and the brain. Many Devonian fishes and primitive tetrapods had a similar intracranial joint, but the function-

ing and biological significance of this structure remain obscure. *Latimeria*'s paired fins move in a fashion that is unlike the coordination seen in most fishes but is exactly the same way as we move our arms and legs, and studies of its nervous system may help trace the beginnings of tetrapod locomotion. The ear of *Latimeria* is also interesting, in that there seem to be sensory areas that are precursors of structures responsible for hearing in air⁴.

These points of scientific interest, along with the general curiosity value of a living fossil, have caused great demand for *Latimeria*. Given an estimated Comoran population of just 500, and the species' low fecundity (it is a live bearer), this has been cause for concern⁵. The possible existence of an Indonesian population may relieve that anxiety, but equally may also mean that other fishing grounds have to be monitored.

For reasons shown in Fig. 1, it is likely that the Indonesian population is quite distinct from that in the Comoros. DNA and protein profiling should provide the unequivocal answer. If the populations are distinct, such work may also allow estimation of the time that they separated. If they are not distinct, then the implication is that the species has a very broad geographical distribution (not unknown in some deep-sea fishes⁶) and that we can expect coelacanths to occur elsewhere in the Indian Ocean, and maybe in the Pacific.

Comoran coelacanths live at about 180 metres, below the 18°C isotherm, and inhabit submarine caves formed through recent volcanic activity⁷. We don't know the exact ecological preference of the Indonesian coelacanths, but these too were caught around a volcanic island known to have submarine caves, and at about the same depth. One theory for the scarcity of *Latimeria* is that the species is competitively inferior and takes refuge in newly established and remote areas⁸. Hitherto, the only known occurrence in the Comoros supported this view. But if more populations are discovered, the less likely it becomes. It could just be that *Latimeria* is a widespread deep-sea fish, which may explain the absence of the coelacanth lineage from the fossil record for so long because deep-sea sediments carrying remains of fossil fish are extremely rare.

The Sulawesi fishermen seem to be well acquainted with the coelacanth; for instance, they have a name for it. It is amazing how the Indonesian occurrence of such a large (adults are up to 1.6 metres in length) and distinctively shaped fish should have escaped the world's attention for so long. The next step will be to estimate the size of the population. We must hope it will be large. At present *Latimeria* is classified under category 'K' ('insufficient known to justify conservation action'). However unfortunate many conservationists think that is, the new discovery



Figure 1 *Latimeria* west and *Latimeria* east. The only natural home of the coelacanth was assumed to be the Comoros archipelago off Madagascar; three specimens caught elsewhere in the western Indian Ocean (capture dates shown) are assumed to be strays swept away from the Comoros by the powerful Mozambique current. Erdmann and colleagues' announcement¹ of finds of coelacanths off Sulawesi, Indonesia, considerably extends the fishes' known geographical range. Given the distance between the Comoros and Sulawesi, the direction of ocean currents and the distinct geological history of the two regions, it is unlikely that the Indonesian specimens are strays.

underlines how little we really know about the coelacanth in particular and oceanic life in general. □

Peter Forey is in the Department of Palaeontology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK.
e-mail: plf@nhm.ac.uk

1. Erdmann, M. V., Caldwell, R. L. & Moosa, M. K. *Nature* **395**, 335 (1998).

2. Smith, J. L. B. *Nature* **143**, 455–456 (1939).
3. Rosen, D. E., Forey, P. L., Gardiner, B. G. & Patterson, C. *Bull. Am. Mus. Nat. Hist.* **167**, 159–276 (1981).
4. Fritzsche, B. *Nature* **327**, 153–154 (1987).
5. Fricke, H., Hissman, K., Schauer, J. & Plante, R. *Nature* **374**, 314 (1995).
6. Nelson, J. S. *Fishes of the World* 3rd edn (Wiley, Chichester, 1994).
7. Fricke, H. & Hissman, K. *Nature* **346**, 323–324 (1990).
8. Thomson, K. S. *Living Fossil: The Story of the Coelacanth* (Hutchinson Radius, London, 1991).

Planet formation

Twin planetary systems in embryo

Alan P. Boss

Size does matter, especially when it comes to radio telescopes. With a maximum extent of 36 km, the Very Large Array (VLA) of 27 radio telescopes yields sharper images at millimetre wavelengths than any other instrument. Rodríguez *et al.*, reporting on page 355 of this issue¹, have used it to produce the most detailed millimetre-wave image to date of the L1551 IRS5 system, confirming the existence of what are probably planet-forming disks around both protostars in this binary system. Their analysis has also thrown up some surprises, and may have considerable implications for the formation of giant planets.

About 500 light years away, Lynds 1551 is an interstellar cloud of molecular gas and dust that is producing young Sun-type stars at a tremendous rate. Embedded within this active cloud is a strong source of infrared radiation, L1551 IRS5, which on its discovery was believed to be a newly formed protostar (less than a million years old) whose optical radiation is degraded to infrared wavelengths through absorption and re-

emission by the intervening dust grains. L1551 IRS5 is celebrated as the young stellar object (YSO) that yielded the first evidence for large-scale outflows of molecular gas, flowing in opposite directions from the young star². These bipolar outflows are now known to be nearly ubiquitous among YSOs. But with the new observations by Rodríguez and his colleagues¹, L1551 IRS5 has even more reason to be celebrated.

The VLA was used in its highest-resolution configuration to map L1551 IRS5 at a wavelength of 7 mm, discerning structures as small as 7 AU (1 AU is the mean Earth–Sun distance). This resolution is about ten times better than previous millimetre-wave observations³. With such a large jump in resolution, one can expect surprises.

The new map (Fig. 1) shows conclusively that L1551 IRS5 is a binary system, with the two protostars about 45 AU apart — not a single object as was once assumed⁴. Furthermore, and even more importantly, there is extended emission from each of the two protostars. They both appear to be circled by protoplanetary disks about 20 AU across. As would be expected, the disks seem to be aligned with each other and in a direction

perpendicular to the small-scale jets that are emitted by the two young stars.

The existence of these disks was suspected from previous observations. But according to the new data, their masses are surprisingly high: about 0.05 solar masses each, roughly three times higher than the previous estimate³. In contrast, disks orbiting somewhat older and more evolved YSOs — the T Tauri stars — are estimated to be around 0.02 solar masses⁵, with much of the mass thought to reside at radii greater than 10 AU (although more material could be hidden in the opaque regions close to the YSO). In any case, these observations imply that young protoplanetary disks can contain much more mass within 10 AU (the radius of Saturn's orbit) than astronomers had been willing to contemplate. As little as 0.01 solar masses might have sufficed to form the planets of our own Solar System, so the L1551 IRS5 disks have more than enough mass to form two complete planetary systems.

Some interesting physics may follow from these measurements. When a disk is massive enough and cold enough, it will be unstable to clumping caused by the self-gravity of the disk⁶. Spiral arms and even gaseous protoplanets might then result⁷, a rapid means of forming giant planets. With roughly 0.05 solar masses, the L1551 IRS5 disks might be able to form gaseous protoplanets in this way in their cool outer regions, although for lower-mass disks this is still less likely than the rival theory, that giant planets form by the 'bottom-up' accretion of smaller bodies.

Giant planets have been detected in orbit around four Sun-type stars that are members of binary systems: Tau Bootis, 16 Cygni B, 55 Rho1 Cancri, and Upsilon Andromedae. Clearly, the presence of a binary star companion does not inhibit the formation of giant planets, at least not for binary separations of several hundred AU or more. But with a separation of only about 45 AU, the L1551 IRS5 binary is close enough⁸ to destabilize planetary orbits of radius greater than about 10 AU. So a giant planet might form rapidly near the edge of one L1551 IRS5 disk, only to be ejected by the gravity of the other star.

That process might explain the peculiar object TMR-1C, thought by Susan Terebey and collaborators⁹ to be a gaseous protoplanet several times as massive as Jupiter that has been ejected from the TMR-1 binary protostar system (Fig. 2). The TMR-1 and L1551 IRS5 systems are quite similar: binary protostars with separations of about 40 AU and 45 AU, respectively, both with the youthfulness of dust-shrouded YSOs. But TMR-1C could turn out to be a background star masquerading as a protoplanet. Spectroscopic observations to determine its true nature are now being made at the Keck Observatory, so we should know soon

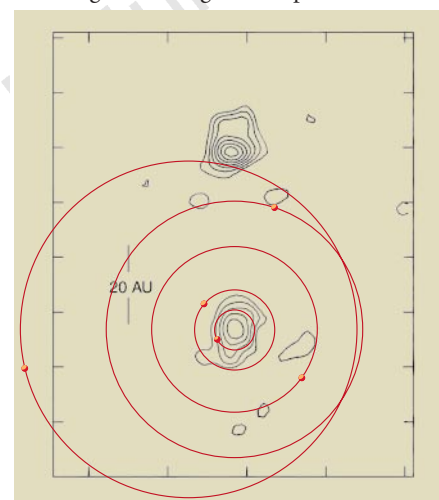


Figure 1 Superimposed solar systems. Contours trace millimetre-wave emission from L1551 IRS5, a binary pair of forming stars. Both are surrounded by disks about as big as the orbit of Saturn, that may be forming planets. The orbits of Jupiter, Saturn, Uranus, Neptune and Pluto are shown centred on one of these stars.



Figure 2 Exiled world? TMR-1C (lower left) might be a giant planet ejected from a system similar to L1551 IRS5.

whether or not TMR-1C is indeed a gaseous protoplanet that might have formed rapidly in a system like L1551 IRS5. □

Alan P. Boss is in the Department of Terrestrial Magnetism, Carnegie Institution of Washington, 5241 Broad Branch Road NW, Washington DC 20015-1305, USA.

e-mail: boss@dtm.ciw.edu

- Rodriguez, L. F. *et al.* *Nature* **395**, 355–357 (1998).
- Snell, R. L., Loren, R. B. & Plambeck, R. L. *Astrophys. J.* **239**, L17–L22 (1980).

- Looney, L. W., Mundy, L. G. & Welch, W. J. *Astrophys. J.* **484**, L157–L160 (1997).
- Shu, F. H., Najita, J., Galli, D., Ostriker, E. & Lizano, S. in *Protostars & Planets III* (eds Levy, E. H. & Lunine, J. I.) 3–45 (Univ. Arizona Press, Tucson, 1993).
- Beckwith, S. V. W., Sargent, A. I., Chini, R. S. & Güsten, R. *Astron. J.* **99**, 924–945 (1990).
- Nelson, A. F., Benz, W., Adams, F. C. & Arnett, D. *Astrophys. J.* **502**, 342–371 (1998).
- Boss, A. P. *Astrophys. J.* **503**, 923–937 (1998).
- Black, D. C. *Astron. J.* **87**, 1333–1337 (1982).
- Terebey, S., Van Buren, D., Padgett, D. L., Hancock, T. & Brundage, M. *Astrophys. J. Lett.* (in the press).

Protein breakdown

Ubiquitous déjà vu

Stefan Jentsch and Helle D. Ulrich

How miserable must life get before one is ready to eat parts of one's own body? Most humans would probably rather starve to death. But nucleated cells are sometimes forced to gather food from their own interior in a process known as autophagy¹, to provide the nutrients needed to keep their vital functions running. However, severe starvation is only one of the factors that can trigger this grisly feast. Autophagy is also one of the main ways for a cell to get rid of obsolete parts, including aged proteins from the cytoplasm and even complete organelles. In busy cells such as those in the liver, for example, the mitochondria that produce chemical energy wear out rapidly and are subsequently digested by autophagy. For obvious reasons, autophagy must be strictly controlled, yet surprisingly few studies have addressed this problem. Now, however, on page 395 of this issue, Mizushima *et al.*² report the

discovery of an intriguing protein-modification pathway that is required for autophagy. This pathway seems to be conserved from yeast to mammals, and may point to a general principle of protein targeting by covalent modification.

Autophagy is the bulk delivery of cytoplasmic material to the cell's central digestion organelles, the lysosomes¹. The process starts with the formation of double-membrane sheets, possibly derived from the endoplasmic reticulum, which envelop a portion of the cell's interior. This newly created vesicle, the autophagosome, seems to travel to the lysosomes along microtubules, which serve as guiding tracks³. On reaching the lysosome, the autophagosome fuses with it. Like our stomach, lysosomes are rather acidic inside, containing a battery of enzymes to break down cellular compounds. In this way, amino acids, lipids, sugars and

nucleotides are recovered and transported back to the cytosol, where they can be re-used by the same cell.

To study the mechanisms of autophagy, Mizushima *et al.*² exploited genetic screens in the yeast *Saccharomyces cerevisiae*. Workers from the same group⁴ had previously isolated 14 autophagy-defective (*apg*) mutants. When several of these genes were cloned, they gave little information about their possible functions. Remarkably, however, Mizushima *et al.* have now discovered a covalent one-to-one complex of two of the gene products, Apg5 and Apg12. Formation of the complex depends on the carboxy-terminal glycine residue of Apg12 and one particular lysine residue in Apg5.

This arrangement brings to mind the interaction between ubiquitin, a protein that labels other proteins for destruction, and its various targets. Ubiquitin becomes covalently attached to a lysine side-chain of the target protein through an isopeptide bond to ubiquitin's carboxy-terminal glycine residue. This conjugation reaction is mediated by a series of enzymes⁵. First, an activating enzyme, E1, hydrolyses ATP and forms a thioester bond between its active-site cysteine and ubiquitin's carboxy terminus. Then a conjugating enzyme, E2, transfers the activated ubiquitin to the target protein. An additional enzyme, E3, may also take part in this thioester cascade.

By analogy to the ubiquitin system, Mizushima *et al.* postulate that there is an enzymatic machinery for the conjugation of Apg12 to Apg5. In fact, the Apg12–Apg5 conjugate was absent in two of the other *apg* mutants that they examined, *apg7* and *apg10*. When they cloned Apg7, the authors found considerable sequence homology to E1 enzymes, including the ATP binding site and the active-site cysteine, supporting the idea that Apg7 may be the enzyme that activates Apg12. Correspondingly, Apg10 might be an E2-like conjugating enzyme. But how might this newly discovered modification system act in autophagy? The localization of the complex at membrane structures suggests a targeting mechanism, possibly during the formation of autophagosomes.

The system discovered by Mizushima *et al.* is the latest in a rapidly growing list of enzymatic pathways for which the ubiquitin system is the prototype (Fig. 1). Earmarking proteins with ubiquitin tags directs them to the 26S proteasome, the main non-lysosomal protease in the cell. Alternatively, some plasma-membrane proteins are directed to the lysosome via the endocytic route⁵, which is unrelated to autophagy. In contrast, conjugation of the ubiquitin-like modifiers Smt3/SUMO-1 (from yeast or mammals, respectively) and Rub1/Nedd8 seems to mediate protein–protein interactions and localization^{6–9}.

What lessons can be learned from these

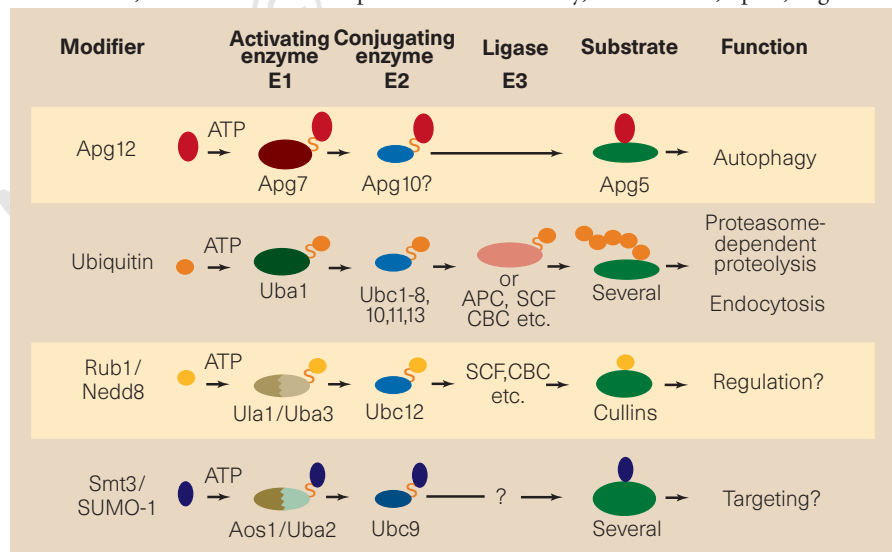


Figure 1 Enzymatic pathways related to the ubiquitin system. Mizushima *et al.*² have found that Apg12 is conjugated to Apg5 in a similar way to the ligation of ubiquitin, or the yeast ubiquitin-like modifiers Rub1 and Smt3 (mammalian Nedd8 and SUMO-1, respectively) to target proteins^{5–9}. The process requires activating (E1) and conjugating (E2) enzymes, and sometimes ligases (E3). E1 (some are heterodimers) and E2 enzymes form thioester (S)-linked intermediates with the modifiers (Apg10 is hypothesized to be an E2). The Apg pathway may not require an E3 because Apg5 seems to be the only substrate². Ubiquitin is often conjugated as a chain, whereas Apg12 and the other modifiers are probably attached as single molecules. APC, anaphase-promoting complex; SCF, SKP-1 cullin-1 F-box protein complex; CBC, cullin-2 elongin B/C complex.

modifier cousins for the control of autophagy? By tagging proteins with targeting signals after they have been translated, timing of the mission can be controlled. For example, conjugation of ubiquitin is tightly regulated and, sometimes, triggered by phosphorylation. Indeed, the Apg12 modifier is phosphorylated². Moreover, another Apg protein, Apg1, is a serine/threonine protein kinase¹⁰ that may be modulated by Apg13, its interacting partner¹¹. Apg1 does not seem to act directly on Apg12 (ref. 2), but it is still an open question whether conjugation of Apg12 is stimulated by Apg1, or by other gene products identified in the pioneering genetic screen⁴. Future experiments will address the consequences of the Apg12 modification and the mechanism of autophagosome formation.

It is gratifying that Apg12 and Apg5 have counterparts in humans^{2,12}, suggesting that the process of autophagy is conserved at the molecular level. Intriguingly, human Apg5 seems to be modified — probably by human Apg12 — on induction of apoptosis¹². This

indicates that a common route may lead to autophagosomes and to the membrane-enclosed vesicles that are typical for programmed cell death. The chances are that the quiet days are over, and autophagy will become a hot field. □

Stefan Jentsch and Helle D. Ulrich are at the Center for Molecular Biology (ZMBH), Heidelberg University, Im Neuenheimer Feld 282, 69120 Heidelberg, Germany.
e-mail: Jentsch@zmbh.uni-heidelberg.de

1. Dunn, W. A. Jr *Trends Cell Biol.* **4**, 139–143 (1994).
2. Mizushima, N. *et al.* *Nature* **395**, 395–398 (1998).
3. Lang, T. *et al.* *EMBO J.* **17**, 3597–3607 (1998).
4. Tsukada, M. & Ohsumi, Y. *FEBS Lett.* **333**, 169–174 (1993).
5. Varshavsky, A. *Trends Biol. Sci.* **22**, 383–387 (1997).
6. Johnson, P. R. & Hochstrasser, M. *Trends Cell Biol.* **7**, 408–413 (1997).
7. Lammer, D. *et al.* *Genes Dev.* **12**, 914–926 (1998).
8. Liakopoulos, D., Doenges, G., Matuschewski, K. & Jentsch, S. *EMBO J.* **17**, 2208–2214 (1998).
9. Osaka, F. *et al.* *Genes Dev.* **12**, 2263–2268 (1998).
10. Matsuura, A., Wada, Y. & Ohsumi, Y. *Gene* **192**, 245–250 (1997).
11. Funakoshi, T., Matsuura, A., Noda, T. & Ohsumi, Y. *Gene* **192**, 207–213 (1997).
12. Hammond, E. M. *et al.* *FEBS Lett.* **425**, 391–395 (1998).

Thermodynamics

Imperfect fluids and inheritance

Glenn H. Fredrickson

The materials that surround us are far from perfect. The polymer molecules that make up the plastics we use to contain our drinks, to mould computer casings and to formulate hair sprays are irregular in both size and shape. Similarly, the latex particles in house paints have a broad distribution of size, surface charge and chemical composition. Even our automotive fuels and lubricants are not simple fluids, but complicated mixtures of linear, branched and aromatic hydrocarbon molecules. When these molecular, chemical and structural irregularities are slight, such ‘complex fluids’ usually have a single phase, which looks homogeneous and is often transparent. Greater irregularity, however, makes a fluid separate into more than one phase, such as the oil and water in salad dressings. Reporting in *Physical Review Letters*, Evans, Fairhurst and Poon have now derived equations that describe how the disorder in complex fluids partitions among the phases¹. Their equations are applicable to fields as diverse as petroleum refining and toiletries, as well as to the design of paints, coatings and plastics.

The fundamental problem can be nicely illustrated in a polymer-science context: ethylene–propylene rubber² (EPR). This commercial synthetic rubber consists of linear polymer chains that have a more or less random sequence of chemically bonded ethylene (E) and propylene (P) monomer units. Because each position along a chain of $N \approx 1,000$ monomer units can be occupied by

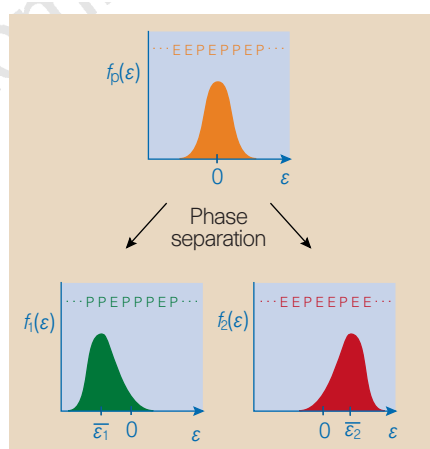


Figure 1 An ethylene–propylene rubber possessing a broad enough distribution of chain composition can separate into two ‘daughter’ phases. ϵ is the fraction of ethylene units (E) on a given polymer chain in excess of the average ethylene fraction. Typical sequences of E and P (propylene) units in chains from the parent and daughter phases are shown above the distribution functions.

either an E or a P monomer, there are roughly $2^N \approx 10^{300}$ possible molecules that can be made by random linkings of the two monomer types. So a macroscopic sample of EPR, containing small multiples of Avogadro’s number of particles (that is, of order 10^{24}), clearly does not contain all of these possible molecules, but it does provide a representative sampling of the composi-

tion distribution of the polymer chains.

As shown in Fig. 1, a ‘parent’ EPR sample with a broad composition distribution can phase separate³ into two or more⁴ ‘daughter’ phases in which chains with more E are spatially separated from chains with more P. Aside from a difference between the mean compositions of the two daughter phases, the full probability distribution functions for composition in these phases can differ considerably from each other, and from the parent distribution.

Evans, Fairhurst and Poon have constructed a mathematical description of such situations that is summarized in the box overleaf. Specifically, they derive an equation (equation (2)) that links the various statistical moments of the parent distribution function — mean, variance, skewness and so on — to other moments of the daughter phase distributions. For example, the daughters’ mean depends on their parent’s variance; their variance on the parent’s skewness. These relationships are valid for all types of complex fluid system, provided that the parent phase is characterized by a narrow distribution function — in other words, as long as the disorder is fairly weak. Evans and colleagues have also made an experimental test of this result for a polydisperse system of latex particles.

In the context of EPR, such connections between the composition distributions of parent and daughter phases could have useful applications. When phase separated, the mechanical performance, optical clarity and other properties of EPR are sensitive to the composition distributions of the daughter phases. So Evans, Fairhurst and Poon have created a tool that should aid the design of new EPR rubbers through manipulations of the parent distribution. In principle, this can be accomplished either by blending different grades of EPR, or by modifying the polymerization process used to manufacture it — changing the catalyst, the temperature or the pressure, for example.

Although I have focused on a very specific application of the new formula, it clearly has much broader implications. For example, phase-separation processes have been used to fractionate emulsions consisting of oil drops in water, or aqueous drops in oil⁵. Equation (2) indicates that it is the skewness (the third moment) of the parent distribution that is responsible for narrowing the drop size distribution of one of the daughter phases. Practically, the consequences are impressive: emulsions with narrow size distributions have already been formulated that actually crystallize into spatially periodic arrays of micrometre-sized drops. These arrays can be fruitfully used in a variety of applications, including the templating of macroporous ceramics for photonic bandgaps, catalyst supports, membranes and filters⁶. The insights of Evans *et al.* may

The mathematics of disorder

An ethylene-propylene rubber can be described by assigning each chain a variable ϵ , the difference between the fraction of ethylene monomers on that particular chain and the sample average. (This is simplified for the purpose of illustration. In practice, more than one random variable ϵ is required to specify the chemical disorder of a random copolymer chain.) So chains with positive ϵ are rich in E units; chains with negative ϵ are rich in P. The chemical disorder that characterizes the full 'parent' sample can be described by a probability distribution function $f_p(\epsilon)$ (Fig. 1), whose n th moment is $\bar{\epsilon}_p^n = \int d\epsilon f_p(\epsilon) \epsilon^n$.

After separating, the two 'daughter' phases will have distinct distribution functions $f_1(\epsilon)$ and $f_2(\epsilon)$ that reflect the locally P-rich and E-rich environments. Evans *et al.*¹ used thermodynamic perturbation theory, assuming a narrow parent distribution, to show that the difference in the means (first moments) of the daughter distributions is proportional to the variance (second moment) of the parent distribution. That is,

$$\bar{\epsilon}_2 - \bar{\epsilon}_1 \propto \bar{\epsilon}_p^2 \quad (1)$$

For a weakly disordered sample, a similar proportionality holds between all higher moments:

$$\bar{\epsilon}_2^n - \bar{\epsilon}_1^n \propto \bar{\epsilon}_p^{n+1} \quad (2)$$

where α is a positive integer. So with $\alpha = 2$, this relates the difference in second moments of the daughter phases to the skewness (third moment) of the parent. The constant of proportionality depends on details of the complex fluid system, but it can be eliminated by forming moment ratios for any two positive integers α and β :

$$\frac{(\bar{\epsilon}_2^\alpha - \bar{\epsilon}_1^\alpha)/(\bar{\epsilon}_2^\beta - \bar{\epsilon}_1^\beta)}{\bar{\epsilon}_p^{\alpha+1}/\bar{\epsilon}_p^{\beta+1}} \quad (3)$$

Amazingly, this universal formula is applicable to any weakly disordered fluid, whatever it is composed of. **G.H.F.**

help to optimize the tedious fractionation process used to prepare such materials and devices.

Evans, Fairhurst and Poon consider situations in which only two phases coexist, but their approach could be straightforwardly extended to any number of phases. Such extensions may prove important in fields such as petroleum refining, where vast numbers of distinct molecular species are present and can partition between several vapour, liquid and solid phases. It is natural to assume that many other areas will be

touched by this development in the physics of disordered materials. □

Glenn H. Fredrickson is in the Department of Chemical Engineering, University of California at Santa Barbara, Santa Barbara, California 93106, USA.

e-mail: ghf@engineering.ucsb.edu

1. Evans, R. M. L., Fairhurst, D. J. & Poon, W. C. K. *Phys. Rev. Lett.* **78**, 1326–1369 (1998).
2. Reisch, M. S. *Chem. Eng. News* 5 August, 10–14 (1996).
3. Scott, R. J. *Polym. Sci.* **9**, 423 (1952).
4. Sollich, P. & Cates, M. E. *Phys. Rev. Lett.* **80**, 1365–1368 (1998).
5. Bibette, J. J. *Colloid Interface Sci.* **147**, 474–478 (1991).
6. Imhof, A. & Pine, D. J. *Nature* **389**, 948–951 (1997).

Developmental biology

Birds of a feather flock together

Mark Peifer

The phenomenon is familiar to anyone who has ever hosted a party. Despite one's best efforts to mix a diverse group of people, the guests rapidly sort themselves out by affinity — in one corner people discuss *Titanic*, in another corner guests talk football, while on the balcony new parents commiserate about sleep deprivation. But sorting by affinity does not apply only to party guests; it is key to the ability of different cells in our body to form organized tissues. Although this has long been suspected on theoretical grounds, and supported by work with cultured cells, reports by Godt and Tepass¹ (on page 387 of this issue) and

González-Reyes and St Johnston² (in *Development*) provide the first compelling evidence that cell sorting by affinity produces organized tissues in the whole animal.

Development is the story of how cells differentiate themselves from one another and, simultaneously, sort out into distinct tissues — a process called morphogenesis. For example, cells of the prospective nervous system must differentiate, segregate from prospective skin cells, and move inside the embryo. Failure to do this correctly results in the birth defect spina bifida. In the 1940s and 1950s Holtfrete and colleagues, in a remarkable series of experiments (see ref. 3, for

example), shed light on this problem. They dissociated tissues such as the prospective nervous system and the future skin into single cells, then mixed them in a dish. The results were dramatic. Although all of the cells could stick to one another, they did so with different degrees of avidity. Cells of the prospective skin and nervous system found their respective compatriots in the cell mix, and sorted out into separate aggregates (Fig. 1). About ten years later, Steinberg⁴ proposed a model for sorting. He suggested that different cells adhere to one another with different strength, and thermodynamic considerations imply that cells stuck together by the strongest glue will associate preferentially.

For a cell to discriminate between neighbours, it must have molecules on its surface that can recognize molecules on adjoining cells. These include proteins of the cadherin family, which are attractive candidates for molecules that could help cells to sort by affinity. First, cadherin extracellular domains project from the surface of one cell and grab the extracellular domains of identical cadherins on neighbouring cells, thereby holding cells together. Second, there are a variety of cadherins, each of which can associate only with an identical cadherin. Thus, if cells expressing two different cadherins are mixed in culture, they sort from one another⁵. Third, different embryonic cells express different cadherins, so prospective neural cells expressing N-cadherin should sort from prospective skin cells that express E-cadherin. Differential expression of cadherins has therefore been suggested as the driving force behind certain morphogenetic events in the whole animal.

Within a single tissue, however, cells often express identical cadherins yet still segregate into distinct cell populations. This is because sorting can also be driven by the expression of different levels of the same cadherin, at least in cultured cells^{6,7}. Cells that express the highest levels of cadherin sort from those expressing lower levels. Again, the party analogy applies — if one tries to avoid cliques by inviting only football fans, they still sort, with the most avid fans glued to the television watching the latest match, and a graded distribution through the rest of the room, with the least avid fans surrounding the beer.

This is where the issue stands in textbooks today — cell sorting in culture can be driven by differences in levels of cadherins. But is this true in the whole animal? It has been difficult to obtain evidence *in vivo*, partly because cadherin-based cell adhesion plays many roles, so defects in one process may obscure defects in later processes. Further, in at least some processes, cell recognition may be mediated by several different adhesion molecules that act in concert, so removing one player

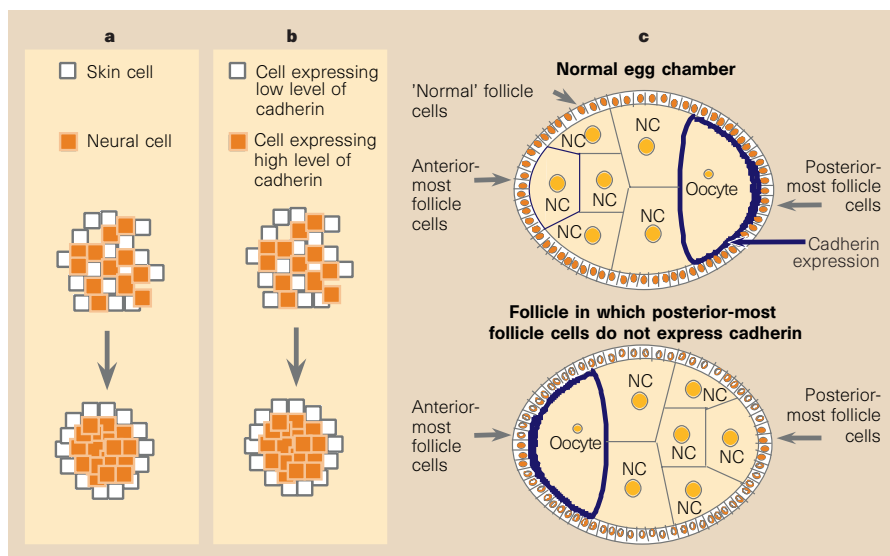


Figure 1 Like poles attract — cell sorting during development. a, *In vitro*, cells can sort themselves according to type, with neural cells preferentially sticking to other neural cells. This is driven by expression of different cadherins. b, Cells can also sort based on the level of cadherin expression, with cells expressing high levels of cadherin preferentially associating. c, Godt and Tepass¹ and González-Reyes and St Johnston² now show that cadherin levels can mediate sorting during development *in vivo*. The oocyte and posterior follicle cells express higher levels of E-cadherin than do nurse cells (NC) and other follicle cells, ensuring that the oocyte is positioned at the posterior pole.

may not disrupt the process.

Now, however, Godt and Tepass¹ and González-Reyes and St Johnston² have the long-sought *in vivo* evidence. Both groups studied follicles of the fruitfly (*Drosophila melanogaster*) ovary, which contain two types of germline cell, the oocyte and its attendant nurse cells. These form a compact package, surrounded by a single layer of epithelial follicle cells. The oocyte is nestled next to the follicle cells at the posterior pole of the follicle, and proper positioning is a key step in setting up anterior–posterior polarity of both the egg and the embryo that develops from it. Thus, we have a small-scale model of morphogenesis — how does the oocyte know to move to the posterior pole?

The solution started with a simple observation. Although all germline and follicle cells express E-cadherin, the oocyte and the posterior follicle cells express higher levels of it than the nurse cells and other follicle cells. This suggested that, as Steinberg and Takeichi⁷ observed in cultured cells, the oocyte might assume a posterior position to maximize its association with the follicle cells that express the highest levels of E-cadherin. (The anterior follicle cells also express increased levels of E-cadherin, but the oocyte may normally never come into contact with them.)

Godt and Tepass, and González-Reyes and St Johnston, manipulated E-cadherin levels to test this hypothesis. First, building on observations by others^{8,9}, both groups found that when E-cadherin was removed from the oocyte and nurse cells, the oocyte was no longer invariably posterior. Instead, it was positioned more or less randomly. They also found that the oocyte adopted a random

position if expression of E-cadherin was lost specifically in the follicle cells. Thus, adhesion between germline and follicle cells is essential for positioning. By itself, this did not prove that cadherins are instructive, so both groups generated 'mosaic' follicles, in which some follicle cells express cadherin and others do not. In almost all cases, the

oocyte moved to be in contact with cadherin-expressing follicle cells (Fig. 1). Even more strikingly, Godt and Tepass found that when the follicle cells expressed different levels of cadherin, the oocyte preferentially moved next to those cells that expressed the highest levels.

The new data provide a direct and dramatic confirmation of Steinberg's theory: first, that morphogenetic movements *in vivo* can be driven by differences in cell adhesion; and second, that differences in the level of a single cell-adhesion molecule are enough to mediate cell sorting in an intact tissue. It is time to rewrite the textbooks, and for cell and developmental biologists to have a party to celebrate the confirmation, *in vivo*, of a venerable theory. □

Mark Peifer is at the Curriculum in Genetics and Molecular Biology and Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-3280, USA. e-mail: peifer@unc.edu

- Godt, D. & Tepass, U. *Nature* **395**, 387–391 (1998).
- González-Reyes, A. & St Johnston, D. *Development* **125**, 3635–3644 (1998).
- Townes, P. L. & Holtfreter, J. *J. Exp. Zool.* **128**, 53–120 (1955).
- Steinberg, M. *Science* **141**, 401–408 (1963).
- Nose, A., Nagafuchi, A. & Takeichi, M. *Cell* **54**, 993–1001 (1988).
- Friedlander, D. R., Mege, R.-M., Cunningham, B. A. & Edelman, G. M. *Proc. Natl Acad. Sci. USA* **86**, 7043–7047 (1989).
- Steinberg, M. S. & Takeichi, M. *Proc. Natl Acad. Sci. USA* **91**, 206–209 (1994).
- Peifer, M., Orsulic, S., Sweeten, D. & Wieschaus, E. *Development* **118**, 1191–1207 (1993).
- White, P., Aberle, H. & Vincent, J.-P. *J. Cell Biol.* **140**, 183–195 (1998).

Isotope astrophysics

Sorting stardust

Sara S. Russell

Certain meteorites, called primitive chondrites, contain microscopic diamonds¹ (Fig. 1, overleaf). These diamonds are of astrophysical as well as mineralogical interest, because they are thought to have formed before the Solar System condensed. In a new experiment — whose results were reported at a meeting in Dublin in July* — two classes of pre-solar diamond have been distinguished, implying that they come from at least two different environments.

We believe that these diamonds may be pre-solar because of their peculiar isotopic composition. The anomalous isotope ratios of trace elements in the crystals imply that they cannot have formed from the well-mixed material that makes up the bulk of the Solar System. Instead, at least some of the diamonds must have formed around other stars. They survived collisions and bombardment by cosmic rays in the interstellar medium and in the contracting disk of the

early Solar System, and remain preserved in some chondrites. Those that fall to Earth and are collected provide an opportunity to study stardust in the laboratory.

So where exactly do the diamonds form? Two possible places are the atmospheres of red-giant stars, and the remnants of supernova explosions. The isotopic composition and chemistry of a given diamond grain should hold clues to the formation site, but because of the tiny size of the diamond crystals (their average diameter is only 1.6 nm), they cannot be isotopically analysed one by one — only their bulk properties can be determined. In this respect, they differ from other identified pre-solar crystals, such as graphite, silicon carbide and corundum, which are large enough to analyse individually, and whose origins are consequently better understood².

For these reasons, little progress has been made in divining the origins of the diamonds. It is not known how many sources contributed to the diamond population, or

*The 61st Meteoritical Society meeting, Dublin, Ireland, 27–31 July 1998.

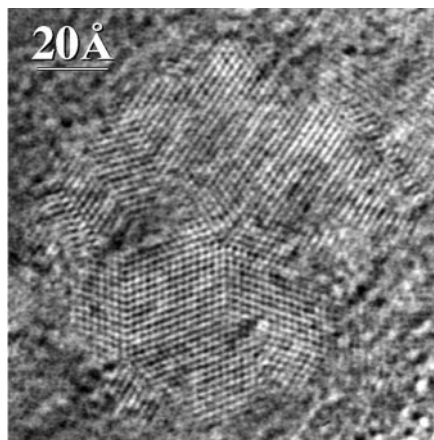


Figure 1 A diamond of about 10^{-17} carats. It is older than the Solar System, and may have condensed in the remnants of a supernova.

what the sources were. Because the carbon-isotope composition of these grains is within the normal Solar-System range, and the isotope anomalies are observed only in trace elements, it is not clear whether all the grains are pre-solar, or just a small fraction of them. It is even possible that the pre-solar material is contained in a different mineral, hidden within the diamonds.

The supernova-origin idea is supported by extremely anomalous xenon-isotope patterns in these samples. Even though this element is so rare that only one diamond in a million contains a xenon atom, it has been shown that the diamonds are highly enriched in the two lightest and also the two heaviest isotopes of xenon. No known single nuclear process can

form such material, but two together could: the lighter xenon isotopes are formed during the 'p' process, and the heavier during the 'r' process, two types of nuclear reaction that are known to happen in supernovae.

The two processes occur in different layers of the exploding star, so in theory heavy- and light-xenon-bearing material should be separable in the laboratory³. Nevertheless, more than 20 years of analysis has not allowed us to untwine these two signatures, indicating either that these diamonds are a single population that formed from isotopically homogeneous xenon, or that the different populations are very well mixed⁴.

The experiment reported in Dublin finally appears to have succeeded in separating the two components (A. Meshik and colleagues, Washington Univ., St Louis). A laser was fired at a collection of diamonds spread out thinly on sapphire plates, and the evolved gases were collected. This technique may owe its success to the fact that some diamonds contain more nitrogen than others⁵. Nitrogen-enriched diamonds tend to be coloured, and so these crystals will absorb more laser light, heat up, and emit their noble gases more efficiently than clear, nitrogen-poor diamonds.

The gases evolved in the latest experiment, presumably mostly from nitrogen-rich diamonds, contained less of the lighter xenon isotopes than bulk diamonds. This result implies that heavier xenon is preferentially located in nitrogen-rich grains and lighter in nitrogen-poor grains, suggesting the existence of two populations of supernova

diamonds. These data confirm our understanding of nucleosynthetic processes occurring in supernovae, and suggest that the expanding shells of the supernovae involved did not mix well, at least until they had cooled enough to allow solid material to condense.

If the result is confirmed, isotopic and chemical analyses of elements other than xenon will be required to investigate the characteristics of the separated components thoroughly. This will constrain the nucleosynthetic processes occurring in specific regions of the supernovae responsible. One outstanding issue that may be resolved by these future experiments is the formation of the heavy-xenon-enriched component. The isotope abundances of this component appear to require condensation of solids just hours after the explosion, contradicting observations that dust only begins to form about a year after the supernova burst⁶.

The new result is grounds for hope that these questions can be answered. Meteoritic diamonds may be forever, but the mysteries surrounding their origins might not last quite so long. □

Sara S. Russell is in the Department of Mineralogy, The Natural History Museum, Cromwell Road, London SW7 5BD, UK.
e-mail: sarr@nhm.ac.uk

1. Lewis, R. S., Tang, M., Wacker, J. F., Anders, E. & Steel, E. *Nature* **326**, 160–162 (1987).
2. Ott, U. *Nature* **364**, 25–33 (1993).
3. Clayton, D. D. *Geochim. Cosmochim. Acta* **40**, 563–565 (1976).
4. Huss, G. R. & Lewis, R. S. *Meteoritics* **29**, 791–810 (1994).
5. Russell, S. S., Arden, J. W. & Pillinger, C. T. *Meteorit. Planet. Sci.* **31**, 343–355 (1996).
6. Richter, S., Ott, U. & Begemann, F. *Nature* **391**, 261–263 (1998).

Behavioural genetics

Worming out social secrets

Some species of the nematode worm (*Caenorhabditis elegans*) are sociable diners, clumping together to share a meal, yet others are more solitary. Why? According to a report by de Bono and Bargmann (*Cell* **94**, 679–689; 1998), these differences can be explained by a change of just one amino acid in a putative neuropeptide receptor.

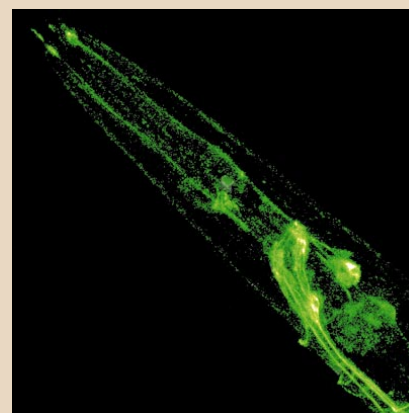
The authors analysed mutations that cause solitary worm strains to form social aggregates. Solitary worms browse alone on bacterial lawns, forming clumps only when food is scarce. But the mutants always swarm together and show other behavioural differences, such as a tendency to burrow into the agar medium.

All of the mutations that caused worms to turn — from solitary to sociable behaviour — mapped to the gene for NPR-1, a seven-transmembrane-domain receptor. The NPR-1 protein closely resembles neuropeptide-Y receptors, which, in humans, are found throughout

the brain. Indeed, de Bono and Bargmann detected *C. elegans* NPR-1 in neurons of the head, ventral cord and preanal ganglion (pictured).

The authors next studied the *npr-1* gene in 15 wild strains of *C. elegans*. Whereas the solitary worms had a valine residue at position 215 of the NPR-1 chain, their sociable counterparts used phenylalanine. This residue is thought to affect the specificity of NPR-1 signalling through guanine-nucleotide-binding (G) proteins, so it may help to transduce the signals that drive the worms to clump.

Could changes in neuropeptide pathways be a widespread mechanism for altering nematode behaviour? Another study by Nelson *et al.* (*Science* **281**, 1686–1690; 1998) suggests that they could. These authors disrupted *flp-1*, a member of the FMRFamide-related neuropeptide gene family in *C. elegans*, and found behavioural defects in the worms, including a lack of



coordination and hyperactivity.

Nematodes probably aggregate because of a mutually attractive stimulus, an as-yet-unknown neuropeptide, that acts through the NPR-1 receptor. Because the worms clump only on bacterial lawns, food is likely to regulate secretion of the neuropeptide — in other words, for worms, as for humans, food is important in social behaviour.

Alison Mitchell

Membrane fusion

SNARE the rod, coil the complex

William I. Weis and Richard H. Scheller

Despite the continuous flux of pro- and anti-parallel membranes must interact for sub-
 cell, out by experiments both *in vitro* and *in vivo*.
 its membrane-bound compartments main- However, the precise function of v- and t-
 tain distinct identities. Their composition is SNARE pairing is not known. Hypotheses
 regulated by the transport of cellular cargo range from this complex acting at an early
 from one organelle to another within mem- step, to determine the specificity of vesicle
 brane-bound vesicles — a region of the docking, to it acting at the last step, to medi-
 donor membrane is pinched off, then it fuses at fusion of the lipid bilayers (Fig. 1).
 with the target membrane and delivers the The cytoplasmic portions of SNARE pro-
 cargo. This process is governed by a highly teins contain several α -helical regions, with
 conserved machinery that includes proteins heptad (seven-amino-acid) repeats in their
 on both the vesicle and target membranes, sequences. The first and fourth residues of
 known respectively as v- and t-SNAREs each repeat are hydrophobic amino acids, so
 (soluble N-ethylmaleimide-sensitive factor one face of the α -helix is hydrophobic. When
 attachment protein receptors). Sutton *et al.*¹ the four helices of the core fusion complex
 (on page 347 of this issue) and Poirier *et al.*² interact, these faces pack against one another
 (in *Nature Structural Biology*) now take a to produce the hydrophobic core of a coiled
 step towards understanding the function coil. The core fusion complex contains three
 of SNARE complexes, by determining the proteins. A neuronal v-SNARE called sy-
 structure of the regions that are important in naptobrevin (also known as VAMP) and a
 forming the core fusion complexes. neuronal t-SNARE known as syntaxin

The expectation that proteins from each contribute a single parallel helix to

the structure^{3,4}. Another neuronal t-SNARE, SNAP-25 (synaptosome-associated protein of relative molecular mass 25,000), contributes two helices, but the orientation of these helices was not known.

Sutton *et al.*¹ and Poirier *et al.*² have now determined the orientation of the heptad-containing portions in SNAP-25. Sutton *et al.* solved the crystal structure of a core synaptic fusion complex containing the heptad-repeat regions of syntaxin-1A, synaptobrevin-II and SNAP-25B, and they find that all of the helices in the core complex are parallel. Poirier *et al.*² came to the same conclusion by replacing residues at the amino- or carboxy-terminal side of the heptads in the SNAREs with cysteine residues, attaching spin labels, and examining their interactions by electron paramagnetic resonance spectroscopy.

Now that we know the structure of the domains that interact in the core complex, what can we say about its function? Figure 1 outlines several possible models. The heptad repeats in each SNARE protein were known to be interrupted by a polar residue⁵, and alignment of t-SNARE sequences reveals that this residue is always a glutamine. The analogous residue in the v-SNAREs is a conserved arginine. Sutton *et al.* show that the three glutamines and one arginine residue interact to form a polar layer in the core of the coiled coil. This polar site may act as a reference point for alignment and assembly of the four helices. Another hypothesis is that the polar layer is a target for the chaperone proteins NEM-sensitive factor (NSF) and α -SNAP (soluble NSF attachment protein; not related to SNAP-25) to initiate disassembly of the complex.

If the core complex determines the specific organization of membrane compartments, two criteria must be met. First, enough different SNAREs must be present and associated with distinct compartments. And second, there must be specific interactions between them. A tally of the known SNARE proteins and their subcellular localization suggests that there are enough distinct proteins to account for the specificity of vesicular transport in mammalian cells. Specific SNARE pairing could be determined principally by the subcellular localization of the interacting proteins, or by preferential high-affinity interactions between particular SNAREs.

Based on the synaptic core complex, other SNARE complexes should contain three helical heptad regions centred on glutamine, and one centred on arginine⁵. Given that v- and t-SNAREs are highly conserved at the interacting heptad positions, what might determine highly specific pairing between particular SNARE combinations? Experiments on other coiled coils indicate that ionic interactions between the residues next to the hydrophobic heptad positions are important for specificity⁶.

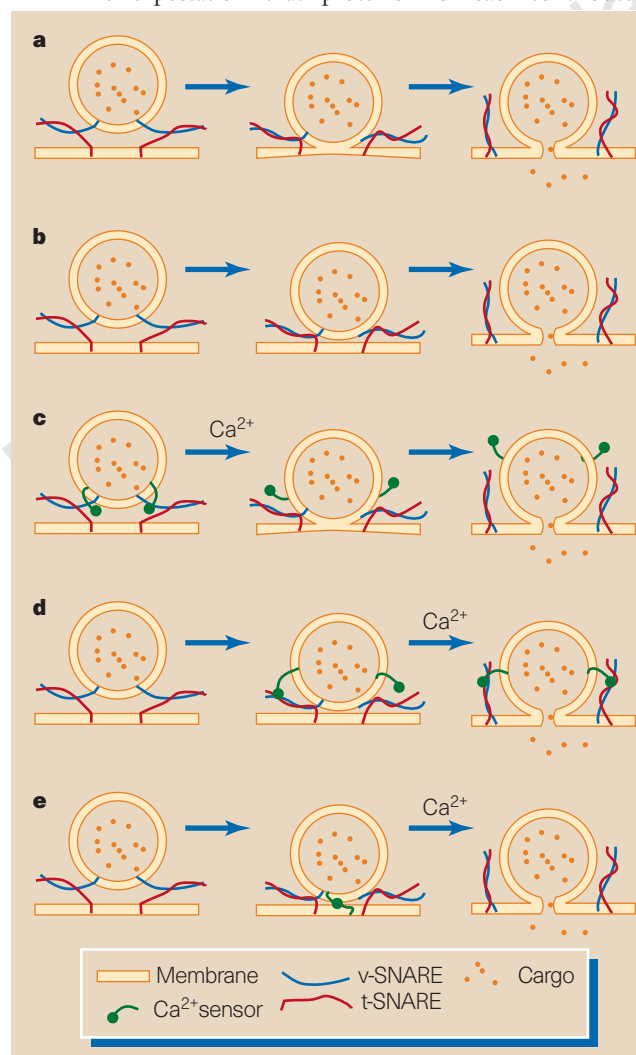


Figure 1 Models of SNARE function. The four-helix coiled coil is represented by a v-SNARE (blue) on the vesicle membrane and a t-SNARE (red), composed of SNAP-25 and syntaxin, on the target membrane.

a, Formation of the core fusion complex drives membrane fusion in such a way that complex formation and membrane fusion cannot be separated. b, As a priming step, formation of the core fusion complex brings membranes in close apposition, but the lipid bilayers do not fuse. Then a second, less well-defined activity fuses the membranes. c, A Ca^{2+} sensor (green) blocks full formation of the complex until Ca^{2+} binding allows fusion to proceed. d, On binding Ca^{2+} , a sensor protein changes conformation. This is transferred to the core complex, resulting in membrane fusion. e, A Ca^{2+} sensor catalyses the final step in membrane fusion without participation of the core complex.

Sutton *et al.* observed a number of salt bridges between syntaxin and SNAP-25 in their core SNARE complex structure, but only two between synaptobrevin and these two t-SNAREs. They speculate that such interactions lend extra stability to pairing between the t-SNAREs. Previous studies have shown that pairing within coiled coils can be prevented by ionic repulsion between the residues next to the hydrophobic heptad positions⁶. Do repulsive interactions at these positions contribute to interaction specificity of the core fusion complex, or might selectivity be determined in a different way from previously characterized coiled coils? Unfortunately, the authors have not listed the ionic interactions or discussed their conservation, so it is difficult to assess the potential contribution of these residues to specificity.

It is possible that formation of SNARE complexes provides affinity and is a driving

force behind bilayer fusion, but that other factors determine the specific organization of membrane compartments. Certainly, the work of Sutton *et al.*¹ and Poirier *et al.*² defines many new questions, and ushers in a new era of high-resolution studies of intracellular membrane fusion. □

William I. Weis is in the Department of Structural Biology, and Richard H. Scheller is in the Department of Molecular and Cellular Physiology and the Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305, USA.

e-mails: weis@fucose.stanford.edu
scheller@cmgm.stanford.edu

1. Sutton, R. B., Fasshauer, D., Jahn, R. & Brunger, A. T. *Nature* **395**, 347–353 (1998).
2. Poirier, M. A. *et al.* *Nature Struct. Biol.* **5**, 765–769 (1998).
3. Hanson, P. I., Roth, R., Morisaki, H., Jahn, R. & Heuser, J. E. *Cell* **90**, 523–535 (1997).
4. Lin, R. C. & Scheller, R. H. *Neuron* **19**, 1087–1094 (1997).
5. Weimbs, T., Mostov, K. E., Low, S. H. & Hofmann, K. A. *Trends Cell Biol.* **8**, 260–262 (1998).
6. O'Shea, E. K., Rutkowski, R. & Kim, P. S. *Cell* **68**, 699–708

has now been questioned⁷. We have only begun to sample the subsurface biosphere, and to understand its organisms and metabolic diversity.

But our understanding of the role of hydrothermal vent systems continues to advance. Such vents are attractive sites for the origin of life, owing to the protection they would have afforded against the comets and asteroids that bombarded the early Earth⁸, and the presence of chemical reducing power in the form of minerals such as iron and nickel sulphides⁹.

Brandes *et al.*¹ have taken advantage of this reducing power, experimentally demonstrating the mineral-catalysed reduction of molecular nitrogen and oxides of nitrogen to ammonia at temperatures between 300 and 800 °C. These results suggest that hydrothermal environments and their surrounding waters would have been the most ammonia-rich environments in the prebiotic world, making them potential oases for early life.

Moreover, vents could have made an important contribution of ammonia to the atmosphere. Provided atmospheric ammonia was partially shielded against ultraviolet photodestruction (by a high-altitude haze, for example), an atmospheric mixing ratio of only 10⁻⁵ of ammonia might have been maintained, providing enough greenhouse warming to keep surface temperatures above freezing¹⁰. (The Sun was appreciably fainter than it is now, and it is uncertain why the Earth was so warm.) So the reducing power of vent environments may have had a global influence on the surface environment, and this sort of coupling between surface and subsurface environments may have been important for the origin of life. ▶

Origins of life

Buried beginnings

Christopher Chyba

Did life begin on Earth near the surface, where the abundant free energy of sunlight could be directly harvested? Or were there instead crucial advantages to be had at subsurface hydrothermal systems? The paper by Brandes *et al.* on page 365 of this issue¹ hints that contributions from both environments might have been required. It increases our appreciation of the opportunities for life offered by crustal and oceanic hydrothermal systems, and raises the possibility that they affected the surface climate.

The possibility that terrestrial life had a subsurface origin will shape the exploration of the Solar System in the next few decades. The science of exobiology has been reborn largely by the discovery of Earth's 'deep, hot biosphere'² — the microbial communities that live beneath the oceans and within the crust — and because of evidence that subsurface liquid water may exist today on Mars and Europa and possibly other bodies³.

If life can only begin on the surface, then Mars might still host a subterranean biosphere. Life could have originated at its surface when conditions were more clement than today, or spread there from Earth. But Europa, whose ocean⁴ (if it exists) has probably been buried beneath kilometres of ice throughout the history of the Solar System, would be barren.

On the other hand, if life can begin in the near or total absence of sunlight, then the putative European ocean is highly intriguing. The US space agency NASA intends to launch an orbiter to Europa in 2003 to test the existence of the ocean and, should it exist,

to begin characterizing it⁵. Later missions should then search for prebiotic and biological signatures.

So can we yet make an educated guess as to which of these ideas is right? Unfortunately, it is not clear whether any of the life under Earth's surface is truly independent of surface photosynthesis. Although it is claimed⁶ that one subterranean methanogen (a bacterium) derives its energy from hydrogen generated from iron-rich basalt, that claim

Earth science

Melting moments

This 1-cm cube of rock provides a persuasive explanation for a common feature of basalts, a type of igneous rock. When basalts form from magma, horizontal sheets of coarser grain segregate out as the magma cools and crystallizes.

As they describe elsewhere in this issue (*Nature* 395, 343–346; 1998), Philpotts *et al.* have examined the process by heating samples of basalt to temperatures of around 1,100 °C, and then quenching and examining serial sections of them. As shown here, the cube retains its shape even when largely melted. The authors think that in the reverse process, crystallization, a three-dimensional crystal network in the basalt starts to maintain its integrity when the surprisingly low level of 25% crystallization is reached. At 35% crystallization, that network becomes strong enough to resist compaction from



overlying material. But in between the two values, it is weak and permeable ('mushy'), and in this interval liquid can be expelled — as is evident from the drop in the photograph. It is this liquid that forms the coarser-grained sheets characteristic of many occurrences of basalt. **Tim Lincoln**

A. R. PHILPOTTS



100 YEARS AGO

Suppose I toss a penny, and let it fall on the table. You will agree that the face of the penny which looks upwards is determined by chance, and that with a symmetrical penny it is an even chance whether the "head" face or the "tail" face lies uppermost. For the moment, that is all one can say about the result. Now compare this with the statements we can make about other moving bodies. You will find it stated, in any almanac, that there will be a total eclipse of the moon on December 27, and that the eclipse will become total at Greenwich at 10.57 p.m.; and I imagine you will all feel sure on reading that statement, that when December 27 comes the eclipse will occur; and it will become total at 10.57 p.m. It will not become total at 10.50 p.m., and it will not wait until 11.0 p.m. You will say, therefore, that eclipses of the moon do not occur by chance. What is the difference between these two events, of which we say that one happens by chance, and the other does not? The difference is simply a difference of degree in our knowledge of the conditions. The laws of motion are as true of moving pence as they are of moving planets...

From *Nature* 22 September 1898.

50 YEARS AGO

That Newton had a just appreciation of the work of Huygens and fully understood it is significant, because Huygens signally failed to comprehend Newton's full achievement, although he realized Newton's greatness as a mathematician and as an experimenter. He criticized Newton's fundamental work on colour because it did not explain the ultimate nature of colour – "Besides, if it should be true that the rays of light, in their original state, were some red, others blue, etc., there would still remain the great difficulty of explaining, by mechanical principles, in what consists this diversity of colours". He did not understand Newton's "But to examine how Colors may be explained hypothetically is beyond my purpose". Huygens himself wrote little about colour, since the problem as he conceived it, to find a mechanical explanation, seemed to him intractable. ... To say this is not to disparage Huygens, whose fundamental achievements make a formidable list.

From *Nature* 25 September 1948.

Other traffic may have run in the opposite direction. For example, it is thermodynamically difficult to form peptides out of individual amino acids in the open ocean, but amino acids have been converted to peptides in experiments that are claimed to model hydrothermal settings¹¹. We don't know whether amino acids themselves can form at vents, but they might have been produced copiously at the Earth's surface¹², and would also have arrived on meteorites¹³. So perhaps they were delivered from the surface to the vents, and only there linked into peptides.

But none of these possibilities should be overplayed; there is still too much we do not know. For example, nitrites and nitrates would have been created by a number of sources on the early Earth, including lightning in clouds of volcanic ash¹⁴, and then reduced to ammonia by Fe²⁺ in the ocean¹⁵. These mechanisms may have produced as much ammonia as the mineral-catalysed reduction identified by Brandes *et al.*¹.

It is intriguing to ask how early terrestrial hydrothermal environments differed from the wet, organic-molecule-rich environments in the parent bodies of carbonaceous chondrite meteorites early in the Solar System's history³. Evidence from mineralogy and organic chemistry makes it clear that the Murchison meteorite, for example, experienced liquid water for the first 10,000 years or so of its history, during which time its amino acids were probably synthesized. Yet there is

no evidence for peptides in Murchison¹⁶, and in general it appears that prebiotic evolution in that object did not proceed beyond simple monomers. Why not? Does this cast doubt on the hydrothermal-origins hypothesis, or was something missing, or the time too short? Within the deep interiors of large asteroids, where liquid water may have persisted for a hundred million years, could prebiotic chemistry have proceeded much further? This is, after all, comparable to the time available⁸ for the origin of life on Earth. □

Christopher Chyba is at the SETI Institute, 2035 Landings Drive, Mountain View, California 94004, and in the Department of Geological and Environmental Sciences, Stanford University, Stanford, California 94305, USA.

e-mail: chyba@seti.org

1. Brandes, J. A. *et al.* *Nature* **395**, 365–367 (1998).
2. Gold, T. *Proc. Natl Acad. Sci. USA* **89**, 6045–6049 (1992).
3. Chyba, C. F. & McDonald, G. D. *Annu. Rev. Earth Planet. Sci.* **23**, 215–249 (1995).
4. Carr, M. H. *et al.* *Nature* **391**, 363–365 (1998).
5. www.jpl.nasa.gov/ice_fire/AO_002.htm
6. Stevens, T. O. & McKinley, J. P. *Science* **270**, 450–453 (1995).
7. Anderson, R. T., Chapell, F. H. & Lovley, D. R. *Science* **281**, 976–977 (1998).
8. Sleep, N. H. *et al.* *Nature* **342**, 139–142 (1989).
9. Holm, N. G. *Orig. Life Evol. Biosph.* **22**, 5–14 (1992).
10. Sagan, C. & Chyba, C. *Science* **276**, 1217–1221 (1997).
11. Huber, C. & Wächtershäuser, G. *Science* **281**, 670–672 (1998).
12. Schlesinger, G. & Miller, S. L. *J. Mol. Evol.* **19**, 376–382 (1983).
13. Chyba, C. & Sagan, C. *Nature* **355**, 125–132 (1992).
14. Navarro-Gonzalez, R., Molina, M. J. & Molina, L. T. *Geophys. Res. Lett.* **25**, 3123–3126 (1998).
15. Summers, D. P. & Chang, S. *Nature* **365**, 630–633 (1993).
16. Cronin, J. R. *Orig. Life Evol. Biosph.* **7**, 343–348 (1976).

Plant domestication

Getting to the roots of tubers

Peter D. Moore

Starch, derived from the roots of tuberous plants, is the staple dietary energy resource for the peoples of many tropical countries. These plants, such as manioc (or cassava, *Manihot esculenta*; Fig. 1) and sweet potato (*Ipomoea batatas*), are widely used, yet little is known about their origins. Undoubtedly they evolved in Central and South America, but beyond that there is little evidence of where or when they were first brought into cultivation. The main problem for archaeologists is the lack of fossils — unlike the grasses, tuberous plants have no persistent parts that survive in sites of former cultivation or food preparation. But an approach that may open up new opportunities for tracing the roots of Neotropical agriculture is now reported in the *Journal of Archaeology*. Dolores Piperno and Irene Holst¹ show that starch grains belonging to some of the plants in question have survived on the surfaces of prehistoric stone tools.

Old World archaeologists tracing the history of the cereals have many advantages over their New World colleagues. Grasses store their starch in tough seeds that are easily

carbonized and preserved in an intact state, allowing them to be identified with precision to species level. The grasses also contain opal phytoliths — silica bodies with distinctive, often angular, shapes — that are relatively inert and survive after organic tissues have decomposed. Moreover, the domesticated cereals produce pollen grains with features that allow them to be separated from other grasses, providing opportunities for the history of cereal-based agriculture to be traced. By contrast, the Neotropical starch-producing root crops, such as manioc, have no identifiable phytoliths, produce few pollen grains and do not carbonize, so it has been difficult to work out their history.

Starch grains vary in size, shape and structure, and it has been claimed that they are distinctive enough to allow species determination². So if starch grains could be found in a datable situation, they could provide evidence for the use of certain plants in the past — including the Neotropical root crops whose history has proved so elusive. The most obvious place to look for such starch grains is on the tools used for grinding and



Figure 1 Ancient roots — heaps of harvested manioc (*Manihot esculenta*) for sale at a market. Piperno and Holst¹ show that starch grains from manioc, and other Neotropical root crops, have survived on the surfaces of ancient stone tools.

preparing food. It has been shown that the stone hunting weapons of prehistoric societies may carry traces of the victim's blood, and that the haemoglobin can be identified to species level³. It is possible that the rough surfaces of grinding tools could, similarly, carry a residue of informative starch grains.

This approach was used by Thomas Loy⁴ in his research into prehistoric remains from the Solomon Islands. He examined stone tools, dating from 27,000 years ago, and found starch grains belonging to *Colocasia esculenta* (taro). This plant has a starch-rich corm (an underground, swollen stem base), and is now widely cultivated in the tropics. Loy's finding provided circumstantial evidence for use of this plant as food in the late Pleistocene and, possibly, even its early domestication. The work also confirmed the potential of starch-grain analysis as a means of tracing the history of early food plants.

Piperno and Holst¹ have now applied similar techniques to investigate manioc and arrowroot (*Maranta arundinacea*) in the humid tropical lowlands of Panama. They examined a number of archaeological sites from which milling stones were available, and extracted starch grains from the crevices in the rough surfaces of these stones using a binocular dissecting microscope and a needle. Not only did they find starch grains, but they also had considerable success in identifying these on the basis of an extensive type collection that they had established. The oldest site was a rock shelter where grinding cobbles dating from the pre-ceramic stage (8,000 to 5,000 years ago) were recovered from sediments in which phytoliths of bottle gourd, squash and maize had been

found. Starch grains that compared well with those of manioc were found on the cobbles, confirming the early use of this plant in the lowland tropics. The authors also found starch grains of maize and arrowroot on these implements.

Much more work needs to be done on the taxonomy of starch grains, particularly those from potential crop species. But these preliminary results seem to establish that a variety of starch-producing plants, including manioc, were being exploited in central Panama about 7,000 years ago. Although we do not know whether these plants were gathered from the wild, or whether they were cultivated, palaeolimnological work in the area indicates some slash-and-burn clearance of the forest at that time. This strengthens the case for the development of agriculture as an explanation for the presence of these plants. The occurrence of other known crop species in association with the manioc and arrowroot gives weight to this argument.

Most important of all is the confirmation that recognizable starch grains can persist over many millennia in the Neotropics — this opens up new opportunities for determining, with precision, how these crop species with an invisible past were domesticated. □

Peter D. Moore is in the Division of Life Sciences, King's College, Campden Hill Road, London W8 7AH, UK.

e-mail: peter.moore@kcl.ac.uk

1. Piperno, D. & Holst, I. J. *Archaeol. Sci.* 25, 765–776 (1998).
2. Banks, W. & Greenwood, C. T. *Starch and its Components* (Edinburgh Univ. Press, 1975).
3. Loy, T. H. *Science* 220, 1269–1271 (1983).
4. Loy, T. H. in *Tropical Archaeobotany* (ed. Hather, J. G.) (Routledge, London, 1994).

Daedalus

A magnetic speaker

The crudest component of a sound system is its loudspeaker. The most flawless recording, perfectly digitized, recovered and ultra-linearly amplified, must still be heard via a clumsy cone of plastic or paper flapping about in time with it. Daedalus now has a better idea. His new 'Paraspeaker' has no moving parts at all.

The Paraspeaker exploits the fact that the oxygen of the air is paramagnetic. So it is attracted into, and compressed by, a magnetic field. Vary that field at an audio frequency, and the pressure must vary in sympathy. This, says Daedalus, is why many transformers produce a steady hum, though they have no moving parts. The air around the transformer is alternately compressed and released by its stray magnetic field, and speaks up at its line frequency. Well — twice that frequency, actually. For, annoyingly, the induced pressure varies not with the field, but the square of the field. This gross nonlinearity doubles fundamental frequencies, and adds other frightful distortions as well.

Daedalus is undismayed. Electronic signal-processing is so fast and precise these days that this and other problems can be sorted out electronically in real time. The Paraspeaker will simply be a large coil of wire set in a hole in a baffle-board; but the electronics driving it will be subtle and complex. The audio signal will go through a square-rooter and programmable signal conditioner while it is still digital. After analog conversion, a dedicated amplifier will feed it to the Paraspeaker.

The Paraspeaker will be a ferociously inductive load; even the best analog amplifier may falter in driving it. The signal conditioner has the job of tweaking the digital signal so as to cancel the aberrations that it will meet downstream. Once programmed to counter the flaws of the following amplifier, it will ensure that a perfect signal reaches the Paraspeaker. So perfect sound will emerge. Its vacant central hole will be an open window on transparent, flawless sound.

The hi-fi community will be entranced. Yet the major benefactors will be the boom-box community, and its victims. For the boom box has but one advantage over a system using headphones: it avoids their discomfort and blanking out of ambient sound. But Paraspeaker earphones, being simply hollow coils, do not obstruct the ears. They can be worn 24 hours a day. The music addict will be able to enjoy his narcotic endlessly, without deafening the rest of us with his damnable loudspeakers.

David Jones

Obituary

Otto Wichterle (1913–98)

Pioneer of biomedical polymers

Otto Wichterle, who died on 18 August at the age of 84, was one of the driving forces behind the emergence of biomedical polymers as a distinct and thriving field of research. Over 40 years ago, Wichterle recognized that polymers to be used in medicine have to be designed and synthesized based on a sound biological rationale; and long before it became well accepted, he took an interdisciplinary approach to hypothesis formulation and problem solving. He will be remembered mainly as the inventor of the soft contact lens, an idea that blossomed into a billion-dollar industry. But his influence is felt on a much broader scale. Research on biomedical polymers has expanded in new directions, and among the results are 'smart' biomaterials, biorecognizable polymers and polymeric carriers of biologically active compounds.

Wichterle was born in Prostějov, in part of the Austro-Hungarian empire that was to become Czechoslovakia, on 17 October 1913. He studied at the Technical University in Prague and obtained his PhD in organic chemistry under Emil Votoček in 1936. After graduation he became an assistant lecturer to Votoček while also studying medicine at Charles University. He was expelled when the German occupation forces closed all Czech universities in 1939. Fortunately, he was able to work in the research institute of the Bata Company in Zlín, where he developed the techniques for the polymerization of lactams. These techniques became the foundation for the Czech polymer industry after the Second World War.

Wichterle returned to the Technical University when the war ended, and in 1949 was appointed professor of macromolecular chemistry and technology. In the late 1950s he created a first-class basic research facility, the Institute of Macromolecular Chemistry of the Czechoslovak Academy of Sciences. Thanks to his leadership, many young scientists (including myself) received an excellent education. Wichterle worked hard and demanded the same from others. I recall his meeting with me and other first-year graduate students in the autumn of 1961, when, among other things, he told us: "Those of you who do not know more about your project than your supervisor within one year should leave at that time". Asked how we could achieve that, he replied: "Your supervisor has about ten



graduate students and many other duties. You would not like to admit that you are ten times less efficient".

His career was interrupted again after the Soviet tanks crushed the 'Prague spring', a short period of liberalization in 1968. He was forced to work alone; nevertheless his patents generated more income than those of all the rest of the institute put together (the staff included more than a hundred holders of PhDs). After the 'velvet revolution' in 1989, which restored democracy in Czechoslovakia, he was elected president of the Academy of Sciences. When he resigned three years later, he was named honorary president.

Wichterle's scientific career started in synthetic organic chemistry. The laboratories of Votoček and Rudolf Lukeš where he studied produced many excellent chemists, including Vladimir Prelog, the Nobel laureate. Wichterle's contributions in this area cover monosaccharide chemistry, the chemistry of alkoxybutadienes and dichlorobutadiene, diene additions, and polymerization of cyclic amides. His main achievement, however, was to open up the field of biomedical polymers to systematic research.

At the beginning of the 1950s, commodity polymers such as poly(methyl methacrylate) and Vinion 'N' (a copolymer of vinyl chloride and acrylonitrile) were beginning to be used for hard contact lenses and vascular grafts, respectively. At that time, Wichterle hypothesized that a polymer with properties close to those of living tissues would be compatible with such tissues, and suitable for use as contact lenses and other implants. Within a year, the first samples of hydrogels were synthesized by cross-linking copolymerization of 2-hydroxyethyl methacrylate and ethylene dimethacrylate. This work (*Nature* 185, 117–118; 1960)

attracted the attention of scientists worldwide.

Wichterle also discovered the so-called spin-casting process for lenses using a children's construction set. In this method, a solution of monomers is converted into a hydrogel lens during a rotation of the polymerization mould. The optical properties of the lens depend on the shape of the rotating form and on the speed of rotation. Numerous modifications of the original hydrogel composition and soft-lens production process were made later, many of them by Wichterle himself. One such development was the technology of preparing soft contact lenses from a dry 'xerogel' by a lathing process, followed by swelling in water.

The transfer of the soft-contact-lens technology to the United States initiated intense interest in hydrophilic materials in general, and hydrogels in particular, to the extent that hydrogel studies were transformed into a whole new field. Among the fruits have been stimulus-sensitive hydrogels, which abruptly change their properties upon the application of an external stimulus, for instance pH, temperature, solvent or electrical field; such changes can be continuous or discontinuous (that is, involve phase transitions). These features give hydrogels potential as actuators or artificial muscles that can convert chemical into mechanical energy.

Hydrogels are also used as carriers for the delivery of drugs, proteins, or as matrices for enzyme or other protein conjugates. Hybrid hydrogels synthesized from synthetic macromolecules and engineered protein domains can be tailored for properties imposed by the structure of the particular domain. Moreover, the field is not restricted to crosslinked polymers. Hydrophilic water-soluble polymers are used for the targeted delivery of anticancer drugs, antisense oligonucleotides and genes.

Wichterle's devotion to science, his energy, creativity and courageous spirit, and his uncompromising stand against oppression of any sort, set an example which formed both the scientific careers and the character of his students and co-workers. All who had the privilege of knowing him will remember him as one of the remarkable scientists of the twentieth century. He was a great teacher, a great scientist and a great man.

Jindřich Kopeček

*Jindřich Kopeček is in the Departments of Bioengineering, and of Pharmaceutics and Pharmaceutical Chemistry, University of Utah, Salt Lake City, Utah 84112, USA.
e-mail: Jindrich.Kopecek@m.cc.utah.edu*

Gray's greyness

What a piece of work is Man! But not in some relentlessly dull anatomy books of the nineteenth century, when 'style' was falling out of fashion. Or perhaps Henry Gray thought the wonders of the body spoke for themselves.

Martin Kemp

The signal visual characteristic that marks the extensive and profound professionalization of the sciences and technologies in the nineteenth century is the progressive dominance of a style of representation that deliberately eschews stylishness.

What I am calling the 'non-style' – a technical mode of illustration in which the dry imparting of information is the sole conscious focus – arrived in different disciplines at different times. Engineering drawing around 1800, especially in France, played a pioneering role. By 1850, there was no branch of institutionalized science that remained untouched, and much twentieth-century illustration is its direct heir.

The transformation was particularly conspicuous in medical illustration. The portrayal of the human body, diseased, deformed or dissected, had always been a fraught business. A heroic mode of illustration displaying elegant figures in brave postures with decorous adornments in gracious settings became the favoured presentation in the erudite humanist picture-books of anatomy, from Andreas Vesalius and Charles Estienne in the Renaissance to Bernhard Siegfried Albinus in the Enlightenment. Such volumes, grand in format, expensive to produce, and often sold by subscription, were hardly the stuff of routine teaching.

The rise of the professional medical school in the nineteenth century did not signal the immediate end of the picture-book,

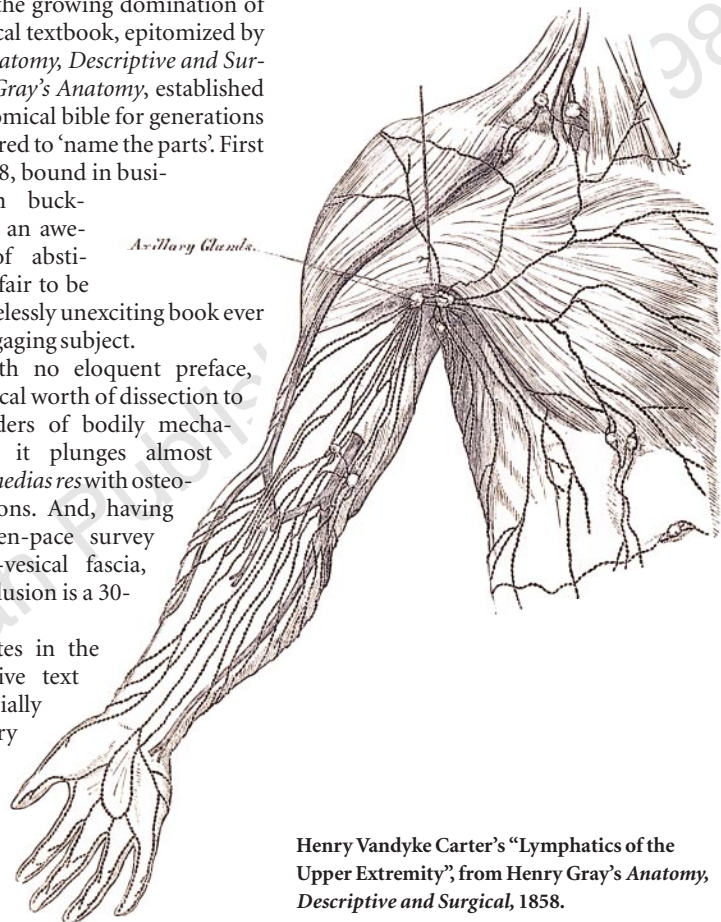
but it did mark the growing domination of the plain, technical textbook, epitomized by Henry Gray's *Anatomy, Descriptive and Surgical*, which, as *Gray's Anatomy*, established itself as the anatomical bible for generations of students required to 'name the parts'. First published in 1858, bound in business-like brown buckram, it achieved an awesome degree of abstinence, and bids fair to be the most remorselessly unexciting book ever written on an engaging subject.

It begins with no eloquent preface, extolling the ethical worth of dissection to reveal the wonders of bodily mechanisms. Instead, it plunges almost immediately *in medias res* with osteological descriptions. And, having finished his even-pace survey with the recto-vesical fascia, Gray's only conclusion is a 30-page index.

The 363 plates in the lucidly descriptive text are by the pictorially named Henry Vandyke Carter, displaying the parts of the body (all parts and no wholes) in sober, matter-of-fact line illustrations. The woodcut technique is used to achieve a single, consistent level of un-seductive description, the register of which is unwavering throughout the book. The plainness is all the more striking when Gray draws directly on the work of his stylish predecessors — debts that are acknowledged in his captions.

A striking case in point is his use of Paolo Mascagni's *Vasorum Lymphaticorum Corporis Humani* (1787). Now famed for his grandiose project for a life-sized human anatomy in three great folios, Mascagni illustrated the lymphatics with vivid pictorial effects in strikingly arranged figures, limbs and organs. Gray took over a few vestiges of Mascagni's pictorial tendencies, but his linear style drains the illustrations of any visual effects that did not serve his rigorously didactic aims.

Whereas Mascagni's large-scale copper engravings demonstrate vessels and glands in three dimensions, modelled in light and shade, Gray's small woodcuts map the course of the lymphatics in a manner that stands in an intermediate position between a natural-



Henry Vandyke Carter's "Lymphatics of the Upper Extremity", from Henry Gray's *Anatomy, Descriptive and Surgical*, 1858.

istic picture and a conventional diagram. Such mapping serves to guide the student in dissection and memorizing, in much the same way that a terrestrial map helps us to plot and memorize our way around a city.

Gray's sterling sobriety survived more or less unscathed through many editions. Photography, sectional anatomy (which developed to a point of considerable sophistication in the second half of the last century), microscopy and, later, X-rays were all long resisted. Gradually, during the course of this century, successive editors have introduced a fresh battery of illustrative techniques, including state-of-the-art microscopy and, inevitably, the latest in computer graphics. Between the brightly coloured covers of the present edition (the 38th), the only consistent visual characteristic that can be discerned is a cacophony of styles and registers of communication. It is most unlikely that Gray would have approved. □

Martin Kemp is in the Department of the History of Art, University of Oxford, 35 Beaumont Street, Oxford OX1 2PG, UK.

e-mail: martin.kemp@trinity.oxford.ac.uk



Henry Gray's *Anatomy, Descriptive and Surgical*: the first edition of 1858.

Indonesian 'king of the sea' discovered

On 30 July 1998, an Indonesian population of coelacanth was discovered. It is apparently the same species as the well-known coelacanth from the Comoran archipelago in the Indian Ocean, *Latimeria chalumnae* Smith.

At sunrise on 30 July, Om Lameh Sonathan and his crew of ten fishermen retrieved a living coelacanth from their deep-set shark gill-net off the young volcanic island of Manado Tua, north Sulawesi, Indonesia. This is almost 10,000 km from the known population of *L. chalumnae* in the Comoros^{1,2}. The Indonesian specimen, which is 124 cm long and weighs 29.2 kg, was observed live by one of us (M.V.E.) for more than 3 hours before the carcass was deep-frozen and tissue samples were collected for molecular analysis.

A preliminary examination of the specimen's external morphology suggests that it is conspecific with *L. chalumnae*, although this must be confirmed by further investigation. The only immediately observable difference from published accounts of *L. chalumnae* is the colour. Previous specimens, from the western Indian Ocean, including the first one observed by Marjorie Courtenay Latimer, are usually described as 'steel blue'^{3,4}, although there are reports of dead coelacanths appearing brown⁵. The live Indonesian specimen was distinctly brown. It shares the same characteristic white mottling pattern as Indian Ocean specimens⁶, but has numerous striking gold flecks over the entire dorsal surface of the body and fins. These are apparently a prismatic effect of light reflecting off the numerous denticles on the scales.

The fish was caught in a shark gill-net of mesh size 6 cm, approximately 150 m in length and 11 m in height, set 3.5 m off the substrate at a depth of 100–150 m for a 12-hour period overnight. The capture site was at the base of a steep volcanic slope known for its complex cave and crevice topography, and appears similar to the habitat reported for the Comoran coelacanths⁷.

This is only the fourth coelacanth reported to be caught in a net, as the Comoran specimens were captured by fishermen hand-lining for the



The Indonesian coelacanth shortly after capture.

oilfish, *Ruvettus pretiosus*⁴. Of the three previous specimens caught outside the Comoros, two were captured in trawl nets, off South Africa³ and Mozambique⁷, and a specimen from Madagascar was caught in a gill-net⁸. Interviews with fishermen throughout north Sulawesi reveal that, although the oilfish is often caught by hand-line in this area, coelacanth (known locally as *raja laut*, or 'king of the sea') are only ever caught using deep gill-nets.

The new specimen is actually the second coelacanth to be reported from north Sulawesi. On 18 September 1997, the wife of M.V.E. saw a strange-looking fish being wheeled in a cart across the fish market in Manado. The fish was immediately recognized as a coelacanth, but we only managed to take some photographs of the fish and briefly interview the fisherman before it was sold. M.V.E. has since been interviewing fishermen in villages throughout the area, as part of a US National Science Foundation international postdoctoral fellowship with the Indonesian Institute of Sciences and with the support of the National Geographic Society. These surveys have identified several

fishermen from north Sulawesi who claim to have captured coelacanths. These interviews, combined with the vast distance from the Comoran archipelago, strongly support the idea that the Indonesian coelacanths are part of an established north Sulawesi population, and not simply 'strays', as has been suggested for the other specimens captured outside the Comoros⁹.

The discovery of an Indonesian coelacanth population has biogeographical and conservation implications. It is unlikely that living coelacanth exist only in two small, highly disjunct populations. Comparison of DNA sequences from tissues of the Indonesian and the western Indian Ocean specimens will reveal the depth of divergence between these two populations.

Further expeditions in Indonesia and to the islands in the vast stretch of Indian Ocean between the Comoros and Indonesia may discover additional populations. This would be welcome news for coelacanth conservation, as the fish is considered highly endangered, in part because of its extremely limited distribution and small population size¹⁰. Nonetheless, the Indonesian government is already considering measures to prevent a repeat of the conservation problems caused by fishing and scientific collection in the Comoros¹¹.

Mark V. Erdmann*†, Roy L. Caldwell*, M. Kasim Moosa†

*Department of Integrative Biology,
University of California, Berkeley,
Berkeley, California 94720, USA

†Indonesian Institute of Sciences,
Jakarta, Indonesia

1. Smith, J. L. B. *Nature* **171**, 99–101 (1953).
2. Fricke, H. & Plante, R. *Naturwissenschaften* **75**, 149–151 (1988).
3. Smith, J. L. B. *Nature* **143**, 455–456 (1939).
4. Balon, E. K., Bruton, M. N. & Fricke, H. *Env. Biol. Fishes* **23**, 241–280 (1988).
5. Millot, J. *Naturaliste Malgache* (suppl.) **1**, 1–26 (1954).
6. Fricke, H. et al. *Env. Biol. Fishes* **32**, 287–300 (1991).
7. Bruton, M. N., Cabral, A. J. P. & Fricke, H. *S. Afr. J. Sci.* **88**, 225–227 (1992).
8. Heemstra, P. C., Freeman, A. L., Wong, H. Y., Hensley, D. A. & Rabesandratana, H. D. *S. Afr. J. Sci.* **92**, 150–151 (1996).
9. Schliwien, U., Fricke, H., Scharl, M., Epplen, J. T. & Pääbo, S. *Nature* **363**, 405 (1993).
10. Fricke, H., Hissmann, K., Schauer, J. & Plante, R. *Nature* **374**, 314 (1995).
11. Forey, P. L. *History of the Coelacanth Fishes* (Chapman and Hall, London, 1998).



Cage structures and nanotubes of NiCl_2

Nanoparticles of layered compounds form hollow cage structures of various types, including nanotubes¹. Here we study nanoparticles of the layered compound NiCl_2 , which can form cage structures, with different numbers of atomic layers, and nanotubes. Because Ni atoms in a layer are coupled ferromagnetically and the interlayer (IIC) coupling is antiferromagnetic, these structures may show different magnetic behaviour according to the parity of closed NiCl_2 layers in the nanoparticles.

NiCl_2 and other three-dimensional transition metal chloride compounds (MCl_2) crystallize in the layered CdCl_2 (space group $R\bar{3}m$) structure. Although within the layers there are strong mixed covalent-ionic bonds between the metal and halide atoms, the layers are held together by weak van der Waals forces. The magnetic moments of the Ni atoms are coupled ferromagnetically in the a - b plane. The moments of adjacent layers are oriented in opposite directions, producing an antiferromagnet with a Néel temperature of 52.3 K (ref. 2). NiCl_2 is a semiconductor with an energy bandgap of 2.4–2.8 eV (ref. 3).

Figure 1 shows transmission electron microscope images of typical closed cage structures, consisting of one closed atomic layer of NiCl_2 with a quasi-spherical (Fig. 1a) or a polyhedral shape (Fig. 1b). In another set of experiments, NiCl_2 cages with three and four layers were obtained.

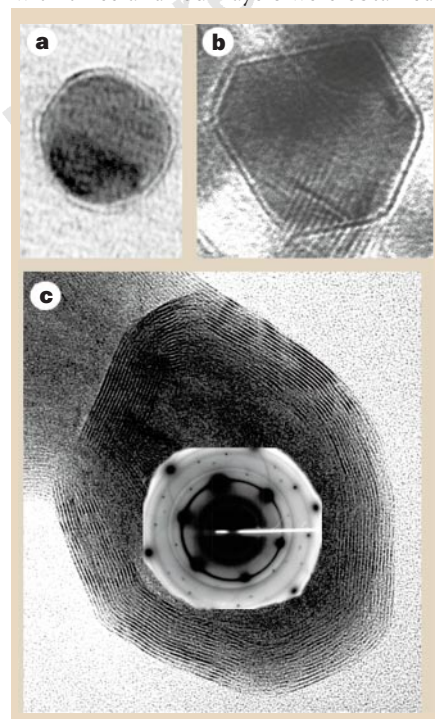
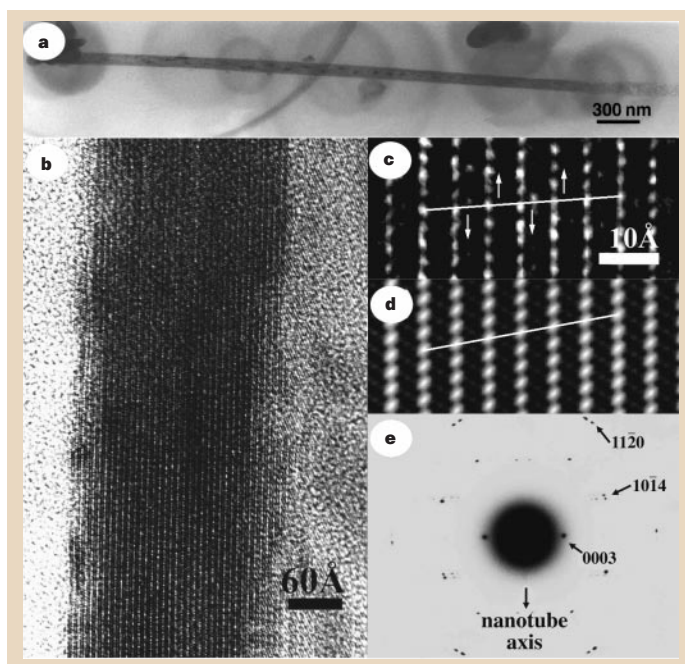


Figure 2 High-resolution transmission electron microscopy (HREM) analysis of NiCl_2 nanotube. **a**, Overall view of the nanotube (low magnification). **b**, Overall lattice image of the wall. The empty central channel is seen on its right. **c**, Higher, digitally filtered magnification of the side-wall region showing the image of eight $c/3$ planes of NiCl_2 . Arrows indicate different displacements of the $c/3$ planes in the nanotube. **d**, HREM image simulation⁴ of the same planes in bulk NiCl_2 .

e, SAED pattern of the nanotube. The $\{10\bar{1}1\}$ and $\{10\bar{1}4\}$ spots come from the planes perpendicular to the curved areas situated between the upper (lower) and side walls of the nanotube.



In both cases, the core of the nested cage structure does not appear to be hollow, although its exact nature is not known. A multiple-layer NiCl_2 hollow cage structure is shown in Fig. 1c. The hexagonal pattern of the selected area electron diffraction (SAED) of the core indicates that the nanoparticle is hollow, which is confirmed by high-resolution transmission electron microscopy (HREM). Energy-dispersive X-ray spectroscopy (EDS) using a probe less than 3 nm across revealed that the nano-particle consisted of nickel and chlorine in an atomic ratio of 1:2.

Figure 2 shows various HREM images of a NiCl_2 nanotube 6 μm long and with a cross-sectional diameter of 70 nm. Figure 2a shows an overall image of the nanotube, whereas Fig. 2b shows a lattice image of the tube wall on one side and the empty channel to its right. An HREM image of NiCl_2 layers of this tube wall and an image simulation⁴ of bulk NiCl_2 in $[01\bar{1}0]$ projection are shown in Fig. 2c and d, respectively. Note the difference in the lattice inclination stemming from the lateral displacements of the $\{0003\}$ NiCl_2 layers (arrows) in the nanotube relative to NiCl_2 bulk.

The corresponding SAED, which has diffraction features typical of a cylindrical nanotube structure⁵, is shown in Fig. 2e. The SAED can be interpreted as a composite diffraction pattern generated by different areas of the nanotube. The two strong $\{0003\}$ spots correspond to diffraction from the planes of the side walls. The six $\{1120\}$ spots are produced by the second prismatic planes perpendicular to the upper and lower walls of the nanotube. It can be deduced that the tube axis has $[1100]$ direction, that is, it is at 30° with respect to the a axis. Some of the spots are doubled, indicating that the structure is chiral, with an angle of 1.8° . EDS indicates that this nanotube contains only nickel and chlorine. The nanotube was stable for a few days.

NiCl_2 is the first of a family of halogen compounds with a layered structure to be shown to form new nanostructures. By combining judicious principles of materials design and sophisticated measurement techniques, it is likely that new physical

Figure 1 Transmission electron micrographs of NiCl_2 nanoparticles with a closed cage structure. **a**, Single-layer quasi-spherical particle (6 nm diameter); **b**, single-layer cage structure with polyhedral shape (10 nm diameter); **c**, many-layer cage structure. Superimposed on the hollow core is its hexagonal diffraction pattern. The distance between the NiCl_2 layers (fringes) is 0.58 nm ($c/3$). Nanoparticles were synthesized by heating hexahydrate in air at 150°C until the powder changes from green to yellow. Dried powder is inserted into the reactor just outside the central hot zone, which is heated slowly to 400°C under N_2 gas flow ($50\text{ cm}^3\text{ min}^{-1}$) for 30 min. The temperature of the hot zone is then raised to 960°C . During heating the gas is slowly switched to Ar ($50\text{ cm}^3\text{ min}^{-1}$). After 30 min the boat is moved to the hot zone. When sufficient loose material has been collected (10–30 min), the temperature is reduced to 200°C while the gas flow is maintained. The sample is desiccated on cooling to room temperature.

phenomena will be discovered in these nanomaterials.

Y. Rosenfeld Hacoheh*, E. Grunbaum*, R. Tenne*, J. Sloan†, J. L. Hutchison†

*Department of Materials and Interfaces, Weizmann Institute, Rehovot 76100, Israel

†Department of Materials, The University of Oxford, South Parks Road, Oxford OX1 3PH, UK
e-mail: cpreshef@weizmann.weizmann.ac.il

1. Tenne, R., Margulis, L., Genut, M. & Hodes, G. *Nature* **360**, 444–445 (1992).
2. Lindgard, P. A., Birgeneau, R. J., Als-Nielsen, J. & Guggenheim, H. J. *J. Phys. C* **8**, 1059–1069 (1975).
3. Ackerman, J., Fouassier, C., Holt, E. M. & Holt, S. L. *Inorg. Chem.* **11**, 3118–3122 (1972).
4. Stadelman, P. A. *Ultramicroscopy* **21**, 131–145 (1987).
5. Whittaker, E. J. W. *Acta Crystallogr.* **9**, 855–862 (1956).

No fractals in fossil extinction statistics

Statistical analyses of the fossil record seek to discover the mechanisms controlling biotic diversity throughout the Earth's history. Solé *et al.*¹ reported that many extinction time series are statistically self-similar, with $1/f$ power spectra, suggesting that extinctions are driven by self-organized criticality or by other scale-free internal dynamics of the biosphere. Here we show that the apparent self-similarity and $1/f$ scaling reported by Solé *et al.* are artefacts of their interpolation methods. Extinction records that are not interpolated show no evidence of fractal scaling.

Self-similarity and $1/f$ scaling would imply that extinction rates are correlated through time. The problem is that similar correlations are created by Solé *et al.*'s interpolation techniques. Their raw data are various extinction metrics² for 77 stratigraphic stages that range up to 34 million years (Myr) in length. They interpolate between these 77 points every million years, creating time series with 570 points, 86% of which are interpolations rather than real data. The Fourier power spectra and root-mean-square amplitude spectra of the interpolated time series follow power laws with exponents close to one (Fig. 1a,b), markedly different from the exponents expected from uncorrelated, 'white noise' data sets, leading Solé *et al.* to infer fractal scaling and attribute it to scale-free dynamics of the biosphere.

But uncorrelated white noise is the wrong null hypothesis in this case, because interpolation introduces points that are correlated with one another. A more appropriate null hypothesis is random numbers substituted for each of the 77 real extinction data, associated with the same stratigraphic stages, and subjected to the same interpolation used in the original analysis. We

evaluated this null hypothesis several thousand times (Fig. 1a,b, red curves), and obtained similar results when we applied the same interpolation and analysis methods to the real extinction data² used by Solé *et al.* (Fig. 1a,b, blue histograms). Random numbers without interpolation yielded the expected 'white noise' spectral exponents (Fig. 1a,b, black curves). These results show that Solé *et al.*'s interpolation methods can generate the fractal scaling they observed.

Solé *et al.* recognized the possibility of interpolation artefacts, but they also found $1/f$ scaling in ammonoid³ and planktonic foram⁴ diversity fluctuations, with no interpolation (Fig. 1c,d). Their results are incorrect; the actual spectra bear no resemblance

to $1/f$ spectra (Fig. 1e,f). Solé *et al.* may have calculated their spectra from time series of taxon counts rather than extinctions or originations. Taxon counts are integrals of the origination–extinction record, so they will be correlated through time and their spectra will be dominated by long-wavelength variations, whatever the dynamics of extinctions and originations.

None of our results suggests fractal scaling in the extinction record or supports recent models of biotic evolution driven by self-organized criticality^{5,6}, which predict that such fractal scaling should occur¹.

The fossil record poses special problems⁷ for conventional spectral methods⁸, which normally require evenly sampled

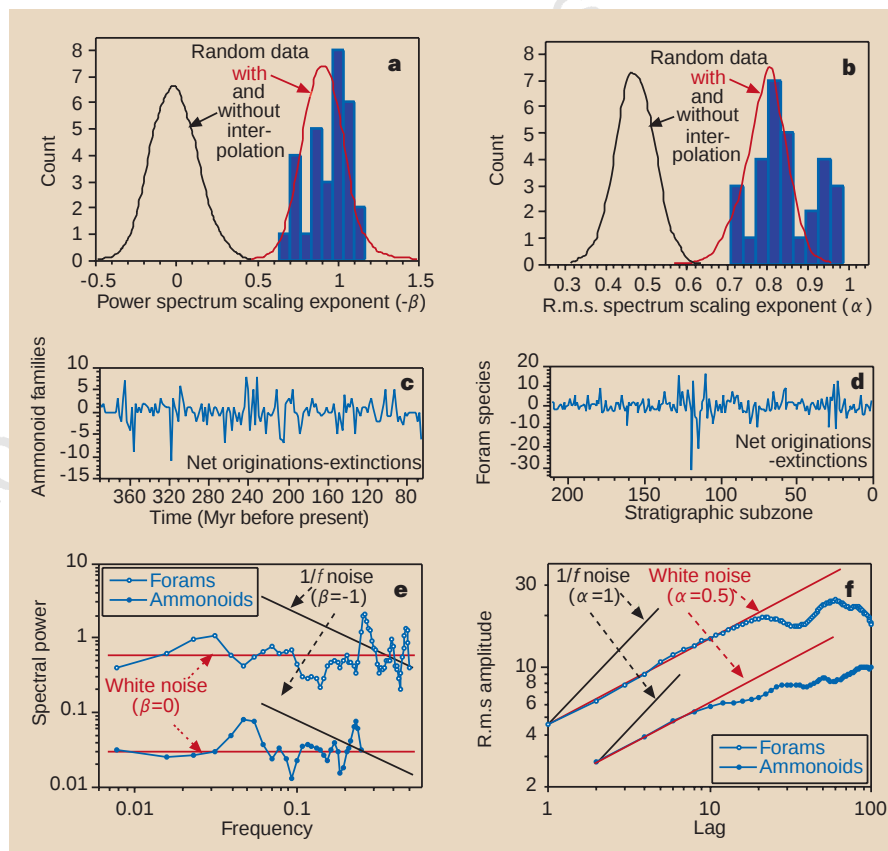


Figure 1 Data analysis. Histograms of fractal scaling exponents for Fourier power spectra (a) and r.m.s. fluctuation spectra (b) for the five extinction measures and six fossil records analysed in ref. 1. Solé *et al.*'s methods yield scaling exponents close to 1.0, suggesting $1/f$ scaling. The same interpolation and analysis procedures yield similar scaling exponents (red curves) when random numbers are substituted for the real extinction data at each stratigraphic stage boundary. Linear and step-function interpolations give similar results, as do random numbers from uniform, gaussian or log-normal distributions. Random uncorrelated data, without interpolation, give 'white noise' spectra (black curves); small deviations from white noise scaling arise because the data series are short. Time series of ammonoid³ and planktonic foram⁴ diversity fluctuations (c,d) yield power spectra (e) and r.m.s. amplitude spectra (f) that deviate markedly from $1/f$ noise (black lines), and are more consistent with white noise (red lines). The foram data⁴ (cited incorrectly in ref. 1) are observations at irregularly spaced stratigraphic subzones, not fixed time intervals. Thus the foram spectra cannot be interpreted conventionally, but we show them because Solé *et al.* report that they obey power laws. Power spectra were calculated from the Fourier transform (with hanning windowing) of the autocorrelation function for lags from 1 to 256 Myr on detrended time series; scaling exponents were estimated by least squares for $0.01 \leq f \leq 0.2 \text{ Myr}^{-1}$. Scaling exponents of r.m.s. spectra were calculated for lags from 1 to 30 Myr. Different procedures yield slightly different results, but in each case the results for the real and random interpolated data series are similar. The exponents reported in Table 1 of ref. 1 are more tightly clustered than the exponent histograms shown in a and b here, owing to errors in algorithms in ref. 1 (R.V. Solé, personal communication).

data from statistically stationary processes where the record is much longer than the longest wavelength of interest. Most fossil time series violate all three constraints. The fossil record is our only possible data source for studying the long-term dynamics of Earth's biotic system, but such studies require carefully designed hypothesis tests.

James W. Kirchner*, Anne Weil†

*Department of Geology and Geophysics,
University of California, Berkeley,
California 94720, USA

e-mail: kirchner@geomorph.berkeley.edu

†Department of Integrative Biology and University
of California Museum of Paleontology,
University of California, Berkeley,
California 94720, USA

1. Solé, R. V., Manrubia, S. C., Benton, M. & Bak, P. *Nature* **388**, 764–767 (1997).
2. Benton, M. J. (ed.) *The Fossil Record 2* (Chapman & Hall, London, 1993).
3. House, M. R. *Phil. Trans. R. Soc. Lond. B* **325**, 307–326 (1989).
4. Patterson, R. T. & Fowler, A. D. *Geology* **24**, 215–218 (1996).
5. Bak, P. & Sneppen, K. *Phys. Rev. Lett.* **71**, 4083–4086 (1993).
6. Solé, R. V. & Manrubia, S. C. *Phys. Rev. E* **54**, R42–R45 (1996).
7. Connor, E. F. in *Patterns and Processes in the History of Life* (eds Raup, D. M. & Jablonski, D.) 119–147 (Springer, Berlin, 1986).
8. Blackman, R. B. & Tukey, J. W. *The Measurement of Power Spectra from the Point of View of Communications Engineering* (Dover, New York, 1958).

Arsenic poisoning of Bangladesh groundwater

In Bangladesh and West Bengal, alluvial Ganges aquifers used for public water supply are polluted with naturally occurring arsenic, which adversely affects the health of millions of people. Here we show that the arsenic derives from the reductive dissolution of arsenic-rich iron oxyhydroxides, which in turn are derived from weathering of base-metal sulphides. This finding means it should now be possible, by sedimentological study of the Ganges alluvial sediments, to guide the placement of new water wells so they will be free of arsenic.

As many as a million water wells drilled into Ganges alluvial deposits in Bangladesh and West Bengal may be contaminated with arsenic^{1–6}. Measured arsenic concentrations^{1–6} reach up to 1,000 µg l⁻¹, which is above the limit set for drinking water in Bangladesh (50 µg l⁻¹) or that recommended by the World Health Organization (10 µg l⁻¹). Consumption of this contaminated water has led to widespread death and disease^{1–6}.

Arsenic has been reported to derive from the oxidation of arsenic-rich pyrite in the aquifer sediments as atmospheric oxygen invades the aquifer in response to a lowering of the water level by abstraction^{4,5}. However, this explanation is not consistent with the following observations³, made on

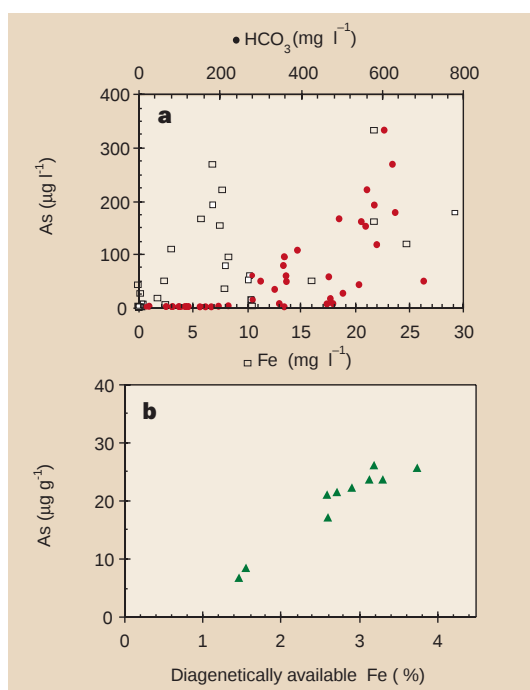


Figure 1 Arsenic chemistry of groundwater and aquifer sediment from Bangladesh. **a**, Arsenic correlates with dissolved iron and with bicarbonate concentrations above 300 mg per litre; the latter is a by-product of iron oxyhydroxide reduction. **b**, Relation of arsenic and iron in aquifer sediments; arsenic concentrations are up to 15 times higher than crustal abundance.

46 wells, typical of those in Bangladesh, that were sampled during May and June of 1997: in oxic (shallow) wells, arsenic concentrations are mostly below 50 µg l⁻¹; in anoxic waters, arsenic concentrations (≤260 µg l⁻¹) correlate with concentrations of dissolved iron (≤29 mg l⁻¹) and bicarbonate (Fig. 1a); and arsenic concentrations increase with depth in wells at Manikganj, Faridpur and Tungipara. These observations suggest that arsenic is released when arsenic-rich iron oxyhydroxides are reduced in anoxic groundwater⁶, a process that solubilizes iron and its adsorbed load and increases bicarbonate concentration. Sedimentary iron oxyhydroxides are known to scavenge arsenic⁷ and, in Ganges aquifer sediments, concentrations of diagenetically available iron (≤3.7%) and arsenic (≤26 p.p.m.) correlate well³ (Fig. 1b).

The arsenic-rich groundwater is mostly restricted to the alluvial aquifers of the Ganges delta^{3,6}. The source of arsenic-rich iron oxyhydroxides must therefore lie in the Ganges source region upstream of Bangladesh. Weathered base-metal deposits are known to occur^{6,8–10} in the Ganges basin (at Bihar, Uttar Pradesh, West Bengal), so weathering of these arsenic-rich base-metal sulphides must have supplied arsenic-rich iron oxyhydroxide to downstream Ganges sediments during Late Pleistocene–Recent times. The arsenic-rich iron oxyhydroxides are now being reduced, causing the present problem. Reduction is driven by concentrations of sedimentary organic matter³ of up to 6%.

A knowledge of the sedimentary architecture and distribution of iron, arsenic and reductant carbon in Ganges alluvial sediments will allow the development of a predictive model to guide future aquifer

development in a way that minimizes arsenic pollution. Furthermore, as dissolved iron is oxidized it precipitates as iron oxyhydroxide, which scavenges arsenic from solution. It follows that simple aeration of anoxic Bangladesh groundwater, followed by settling, should remove a considerable amount of arsenic from solution. This simple treatment could be performed on a household or village scale. Although the disposal of the arsenic-rich iron oxyhydroxides would require special arrangement, this would be preferable to either the widespread poisoning that now exists or a return to the use of contaminated surface water for public consumption.

Ross Nickson*, John McArthur*, William Burgess*, Kazi Matin Ahmed†, Peter Ravenscroft‡, Mizanur Rahman§

*Geological Sciences,
University College London,
Gower Street, London WC1E 6BT, UK

†Department of Geology,
University of Dhaka,
Dhaka 1000, Bangladesh

‡Mott MacDonald International Ltd,
122 Gulshan Avenue, Dhaka-1212, Bangladesh
§GWCII, Bangladesh Water Development Board,
72 Green Road, Dhaka 1207, Bangladesh

1. Pearce, F. *The Guardian* (19 & 26 February, 1997).
2. Dhar, R. K. *et al. Curr. Sci.* **73**, 48–59 (1997).
3. Nickson, R. Thesis, University College London (1997).
4. Das, D., Basu, G., Chowdhury, T. R. & Chakraborty, D. in *Proc. Int. Conf. Arsenic in Groundwater* (Calcutta, 1995).
5. Saha, A. K. & Chakraborty, C. in *Proc. Int. Conf. Arsenic in Groundwater* (Calcutta, 1995).
6. Bhattacharya, P., Chattarjee, D. & Jacks, G. *Water Res. Dev.* **13**, 79–92 (1997).
7. Mok, W. M. & Wai, C. M. *Arsenic in the Environment* (ed. Nriagu, J. O.) 99–117 (Wiley, New York, 1994).
8. Ghosh, S. & De, S. *Ind. J. Earth Sci.* **22**, 183–189 (1995).
9. Bannerjee, D. K. *Mineral Resources of India* (World Press, Calcutta, 1992).
10. Wadia, D. N. *Geology of India*, 4th edn (Tata, 1975).

Beware scientists writing history

The Atom in the History of Human Thought

by Bernard Pullman

Oxford University Press: 1998. 416pp. \$30, £25

J. L. Heilbron

Twenty-five centuries ago, Democritus delivered the Word: "There is nothing but Atoms and the Void." Few listened, there being no evidence for it, except, in time, the opposition of the Roman Catholic Church, often a good guide to truth in physics. Two hundred years ago, the first unequivocal signs appeared, in the form of John Dalton's theory of the atom. Witnesses to the truth sprang up abundantly; and the Word would have prevailed five score and more years ago but for some pernicious, prejudiced, pestiferous Positivists. They have now joined their ancestors. The Word has triumphed, most wondrously: for the Word Triumphant, the internally structured atom ruled by uncertainty, is not the Word of Democritus, or anything like it, or even utterable in words. That is all the more reason to exalt the Witnesses, and to condemn their critics.

Let it not be supposed that this paraphrase of *The Atom in the History of Human Thought* contains the least element of caricature. The parallel between Bernard Pullman's book and bad church history runs deep. The author, formerly professor of quantum chemistry at the Sorbonne, abandoned the standards of his discipline and of all scholarly inquiry when he took up his historical enterprise. He did not inform himself sufficiently about his subject or check the sources on which he relied. The method works no better in history than it does in chemistry. Scientists who enter seriously into what is for them a new field of science learn the subject before writing about it. Scientists who write history without mastering the main relevant facts in plain sight show that they do not take their historical work seriously. Why should anyone else?

Two blunders, which are the more telling because the correctives were especially easily accessible in the historical literature, will indicate the trouble. In his eagerness to prove the perniciousness of positivism, Pullman "quotes" (the reason for the inverted commas will appear shortly) a criticism of Ernst Mach published by Albert Einstein in 1922. But it is well known that Einstein fancied himself a follower of the positivistic school when he invented the theory of relativity around 1905. After obtaining his PhD, he wanted to work under Wilhelm Ostwald, a minor fiend in Pullman's anti-pantheon, and he sent his published work on relativity to the arch-fiend Mach, as an indication of



Atom theorists Democritus (above, by Antoine Coypel, 1692) and John Dalton.

progress consistent with Mach's epistemology. The occasion for Einstein's later criticism was the posthumous publication of Mach's *Principles of Optics*, which carried a preface rejecting the theory of relativity. No doubt Einstein had come to see that his work did not align perfectly with Mach's prescriptions. But that does not change the fact that Einstein gained insight and guidance from positivism during the most creative period in his life.

Pullman's blunder here is typical of writers who can see no creative historical role for outmoded ideas or for points of view they oppose. Another example of the same type is his condemnation of René Descartes as an enemy of the Word. To be sure, Descartes was no Democritus; but his system was by far the strongest force that the mechanical philoso-



phy of the seventeenth century mounted against the Aristotelian philosophy of the schools. Cartesianism made common cause with atomism for the most important position urged by both: that our minds and senses create almost all the qualities that the school philosophy attributed to the material world.

The second sort of blunder is committed by many scientists who write about history.

They rely on one another. Their own technique should alert them to the danger of doing so.

Pullman makes the double mistake of trusting Werner Heisenberg's apologetic and biographical writings. These are full of quotations that Heisenberg made up. They represent what he thought people should have said, not what they said in fact. Pullman puts these Thucydidean speeches into the mouths of his actors as if he were quoting them verbatim. "At a conference held in Göttingen in 1922, he [Niels Bohr] made the following comments about the genesis of his own model." There follows a long quotation from Heisenberg that jumbles together the sorts of things that Bohr might have said at different times and places in words he would not have used. Pullman "quotes" Wolfgang Pauli, Erwin Schrödinger and, of course, Einstein, in the same uncritical way, from the same unhistorical source. We have history as Heisenberg wanted us to have it. We should reject it.

There are many quotations in Pullman's book that might be informative. But how is the ordinary reader to know which are properly taken, in context, as attributed, and which are not? One reason that scientists are so concerned to be apt and accurate in their own work is that the detection of a few howling errors casts doubt on a body of work that might otherwise be sound. The same is true for historical writing. □

J. L. Heilbron is at Worcester College, Oxford OX1 2HB, UK.

It's written in stone, but can we trust it?

The Adequacy of the Fossil Record

edited by Stephen K. Donovan and Christopher R. C. Paul

Wiley: 1998. 300pp. £50, \$85

Andrew B. Smith

In *The Adequacy of the Fossil Record*, Stephen Donovan and Christopher Paul have assembled a series of reviews that bring the question of the fossil record's quality into sharp focus. It is undeniable that the fossil record is incomplete and can never be perfect, but how serious a handicap is this? After all, no science is founded on complete knowledge, and scientists have always had to make the best use of what information lies to hand.

The key question, then, is not about completeness, but whether available data are adequate to justify the hypotheses and conclusions being drawn from them. In particular, are there confounding factors that might create misleading patterns among any set of observations made?

Unfortunately, palaeontology suffers much more from distorting biases than the life sciences. Although both disciplines are



Fragments from a broken record, clockwise from top left: *Dapedius obicularis*, an enamel-scaled fish from the Lower Lias in Dorset; the Ordovician starfish *Platanaster ordovicus*; a slabful of Silurian brachiopods, *Camarotoechia*; the Jurassic dragonfly *Urogomphus eximus*, from the Kimmeridgian Lithographic Stone in Bavaria.

potentially affected by biases in the way samples are collected, palaeontology is uniquely afflicted by a variety of preservation biases that have drastically reduced and distorted the pool of data that can be sampled. Palaeontologists must therefore be particularly vigilant about the quality of their data before accepting patterns in the fossil record as genuine.

Is sampling adequate? Clearly palaeontologists need to know whether enough material has been collected to be confident that the patterns that emerge are representative and reliable. Sampling biases in data gathering are probably the easiest to take into

account. Many of the problems are identical to those facing ecologists and can be compensated for with knowledge of the sampling regime employed. There are some uniquely palaeontological problems, such as how to estimate error bars associated with stratigraphic ranges, but the appropriate methodologies for these are currently being worked out. All help to prevent over-interpretation of results.

Is there adequate fidelity of preservation? Under special conditions the preservation of fossils can be extremely good, even down to the sub-cellular level in some amber-entombed insects and phosphatized fishes.

Unfortunately, where high-quality preservation is really needed, details are often sadly lacking and interpretation becomes ambiguous, as with the Ediacaran biota. The palaeontological literature has been bedevilled by erroneous reconstructions based on misinterpretation of far from ideal material. A more pervasive bias arises from variation in the structure and chemical composition of skeletons. This directly influences the preservation potential of fossil taxa, but, so long as this is recognized, serious problems can be avoided.

Are there biases in the sedimentary record preserved at outcrop? Sequence stratigraphy has emphasized how sea-level change controls sedimentary packaging, and it is becoming evident that it can also influence our perception of diversity through time. Much more attention needs to be paid to the biases imposed by the depositional architecture within sedimentary basins.

Finally, are our samples of fossiliferous outcrops sufficient to provide an undistorted picture of the history of life through time? Any long-term change in the type of habitat that ends up being preserved might have a significant effect on perceived diversity. For example, whereas the marine shelly macrofauna of the north-west European Turonian chalk facies is now thoroughly documented, contemporary marginal inshore facies are almost non-existent. Although these must have been deposited around emergent land masses, they simply have not survived and their biotas have been largely lost. Such large-scale bias in facies preservation is hard to compensate for without access to phylogenetic information and must presumably affect regional and global diversity curves.

Statistical methods recently devised to assess the completeness of the fossil record from observed distributions suggest that the record is surprisingly good for some taxa. Furthermore, analyses of the growth in palaeontological knowledge have demonstrated that our view of the fossil record is mature and relatively stable. There are unlikely to be many major surprise discoveries that significantly extend the stratigraphic ranges of well studied groups. However, both these approaches indicate only that our knowledge of fossil deposits that have survived and can potentially be sampled at outcrop is becoming rather good. They tell us little about how representative or systematically biased that record is. Stability does not necessarily equate with accuracy.

In *The Adequacy of the Fossil Record*, many of the problems outlined above are dealt with in a clear and stimulating way, making this volume required reading for all those trying to make sense of the fossil record. In the past, palaeontologists have been too uncritical of the data that they can obtain and too willing to read the record at face value. Much can be done with the fossil record, but progress will

not be made unless palaeontologists start to collect data in a more rigorous fashion and face up to the real problems of bias that beset their science. This book represents a major step in the right direction. □

Andrew B. Smith is in the Department of Palaeontology, The Natural History Museum, London SW7 5BD, UK.

Rambling along the byways of probability

Statistical Visions in Time: A History of Time Series Analysis 1662–1938

by Judy L. Klein

Cambridge University Press: 1998. 314pp.

£45, \$64.95

A. W. F. Edwards

In 1826, Samuel Heinrich Schwabe of Dessau in Germany started a systematic series of observations on sunspots, recording each group as it traversed the Sun's disc. Twelve years later, he published his counts, without comment, in *Astronomische Nachrichten*. But in 1843, when he published a continuation, he remarked that there seemed to be a periodicity of about ten years in the counts, most strikingly evident in his record of the number of days in each year without spots.

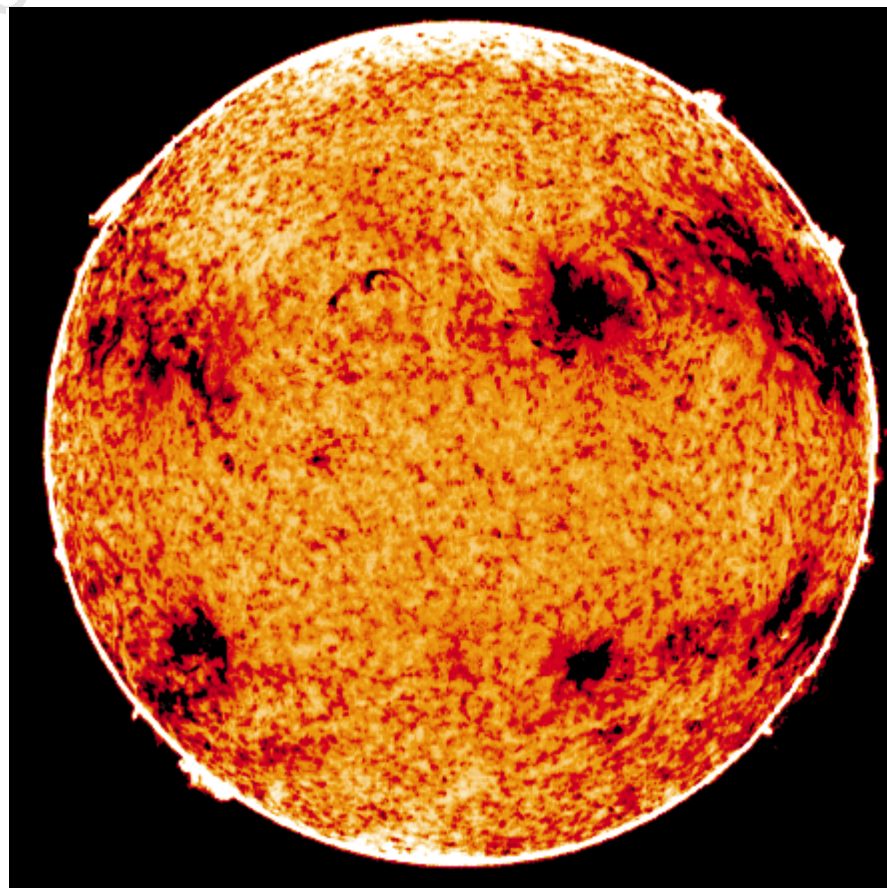
Thus was inaugurated the most famous of all 'time series' (a series of values of a vari-

able taken in successive periods of time): the eleven-year cycle of solar activity, the 'sunspot cycle'. Rudolf Wolf, the first director of the Swiss Federal Observatory at Zurich, was able to estimate the epochs of maximum and minimum sunspot activity back to 1610 by searching earlier records, and to create an index of solar activity — the Wolf sunspot number — from 1750 onwards.

The first example of correlations between time series soon followed, when a number of workers simultaneously, in 1851–52, remarked on the high degree of association between sunspot activity and the magnitude of variations in the elements of the Earth's magnetic field.

In 1879, George Gabriel Stokes suggested a method of searching for periodicities in the sunspot number which Arthur Schuster, in 1898, developed into the periodogram of harmonic analysis, itself later developed into spectral analysis. By 1927, G. U. Yule had come to appreciate that harmonic analysis was inadequate for the analysis of economic time series, and introduced the notion of autoregression for modelling them. In connection with the sunspot cycle, Yule then suggested that the periodicity was disturbed in a way which would now be described in terms of chaos theory.

The analysis of time series has become an important branch of statistics, especially in economics. But in *Statistical Visions in Time* the reader is hard put to detect the backbone



Spot-on: an image taken through a neutral helium filter at the US National Solar Observatory, Arizona.

NASA

of the historical development of the theory. Endless excursions into topics in the history of statistics and probability (which have been better described by other recent writers) obscure the main theme, while the long quotations and cramped footnotes in a minuscule type further detract from a coherent account.

The author seems to have become fascinated by the Galton and Pearson archives in University College London, resulting in an overload of peripheral material. Thus the large number of diagrams oddly includes both Alfred Russel Wallace's sketch of an irregular frequency curve, and the distribution of head breadth in W. F. R. Weldon's crabs, while a whole page is devoted to "Bowditch's composite photograph of 287 American female college students" (all these from the Galton and Pearson archives). But the series of sunspot numbers is nowhere to be found, either tabulated or depicted, and poor Schwabe does not even get a mention.

The history of probability and statistics has been well-served by recent writers, but there is still a gap to be filled around the turn of the century. We lack a biography of Karl Pearson, while W. Stanley Jevons, F. Y. Edgeworth and Yule are only now beginning to emerge from the shadows. *Statistical Visions in Time* contains much relevant material for this period, but not in a coherent form. Indeed, the book is rather like a time series itself: the underlying message is difficult to interpret because of the overlay of random noise. A shorter series with less noise would have been more informative. □

A. W. F. Edwards is at Gonville and Caius College, University of Cambridge, Cambridge CB2 1TA, UK.

Life is like a game of chess

Emergence: From Chaos to Order

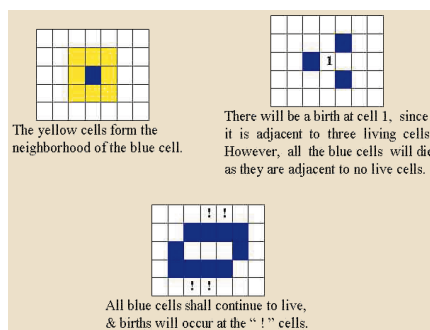
by John Holland

Addison-Wesley: 1998. 258pp. \$25

Hanspeter Mallot

Nature is full of wonders. Trees and flowers grow from inconspicuous seeds, simple physical laws underlie complex phenomena such as climate change or the movement of celestial bodies, and even man-made games such as chess or Go keep generating unsolved problems. Do such phenomena form a well-defined class or is it only the amazement they generate in the observer that groups them together? John Holland takes the view that there is a common principle or rule underlying phenomena where "much comes from little"; he adopts the term "emergence" to characterize this principle.

In his latest book, Holland examines a number of arenas of emergent behaviour,



Game of Life: the rules.

including board games (chess and checkers), machine learning, cellular automata and neural networks. But what exactly is emergence? Holland is not very explicit about this question, leaving the reader with hints such as "much coming from little" or "the whole is more than the sum of its parts". In order to make that into a definition of any value, one needs to say something about the measures used to judge much and little or more and less. No such attempt is made in the book, and, in fact, I doubt that a convincing definition (except for the observer's amazement) can be given. Instead, Holland uses the term "emergence" in an associative and exploratory way.

By and large, emergence seems to be something that arises in the context of simple theories, games or mathematical axiom systems, that is, in models of the world, rather than in the world itself. In this sense, unanticipated predictions may emerge from a physical theory or certain chess strategies may emerge from the rules of the game. One could call this a logical notion of emergence since it is confined to the realm of theory. A related, epistemological notion of emergence views it as the inverse of reduction. This again is a model-immanent view that is not much different from surprise or amazement.

Holland also uses the term in other, more puzzling ways, where "emergence" seems to refer to the causes of real world phenomena. One view is psychological in nature: learning and creativity are called emergent since they lead to behavioural competences or insights that did not exist before emergence occurred. In its most equivocal version, the term occurs on the first page of the book, where we read that "we will not understand life and living organisms until we understand emergence". Here, emergence appears to be a natural phenomenon that causes the origin of life. It is clearly this connotation of a vital force that gives the term its appeal, even though Holland tries to avoid such associations in later parts of the book.

What do we gain by adopting the term "emergence" as a scientific notion? Do we really learn something about the origin of life, say, if we understand machine learning? Of course, there is always the possibility of

cross-fertilizations between far distant fields of science, but this is not the kind of insight the author seems to promise. What he is interested in is a general theory of complexity of which emergence would be a part.

The book is written for a generally educated reader. The examples selected for the illustration of emergent behaviour are very interesting. (In fact, I think that this interest is what measures "much and little" in Holland's explication of emergence.) Holland uses John Conway's Game of Life, computer programs learning board games, and neural networks to illustrate what he has identified as building blocks of emergence: the interaction of simple mechanisms. While making this general point, there are a number of inaccuracies in the details that limit the value of Holland's presentation. To mention just one example, the proposed mechanism of visual pattern recognition based on Hebbian learning and saccadic eye movements is biologically highly implausible, not least for the exceedingly long time required to recognize even simple triangles.

Towards the end of the book, Holland states: "at the broadest level, this book's thesis has been that models and model building underlie much, perhaps most, human intellectual endeavor". While this statement is too broad to be challenged, it makes an important point about the book: Holland is not so much interested in the analysis of natural systems, but in the construction of artificial ones. In this context, the problem of emergence may actually be a genuine one. The empirical part of natural science is reductionist or top-down, whereas in the construction of systems (and theories) one proceeds bottom-up. So if we view emergence as the inverse of reduction, here it may have its valid place. □

Hanspeter Mallot is in the Max Planck Institute for Biological Cybernetics, Spemannstr. 38, D-72076 Tübingen, Germany.

correction

In the 27 August issue (*Nature* **394**, 844–845; 1998) we published a review of *The Story of Interferon* by Kari Cantell. The review repeated allegations made in the book about the late Frank Rauscher, a former senior vice-president of research (not president, as stated) of the American Cancer Society. The allegations concern the purchase by the ACS of interferon from a US company, Life Sciences, in preference to the Finnish Red Cross Blood Transfusion Service. *Nature* and Dr Cantell accept that Dr Rauscher was not one of the founders of Life Sciences and that neither Dr Rauscher nor anyone else at the ACS had any financial interest in the company, or stood to gain financially from the transaction. We apologize to the ACS and to Frank Rauscher's family for this misrepresentation.

Role of plagioclase crystal chains in the differentiation of partly crystallized basaltic magma

Anthony R. Philpotts, Jianyang Shi & Caroline Brustman

Department of Geology and Geophysics, University of Connecticut, Storrs, Connecticut 06269, USA

Melting experiments on samples of basaltic rock from a thick lava flow reveal that when this lava originally began to crystallize, feldspar crystals linked together to form a continuous three-dimensional network of chains when the lava was no more than 25% crystallized. Formation of this network has profound implications for the behaviour and differentiation of basaltic magma. Much of the compositional variation of igneous rocks results from the separation of liquid from crystals, a process that is dramatically affected according to whether crystals occur separately or are linked together in networks.

Many thick flood-basalt flows^{1–7} and lava lakes^{8–11} contain horizontal sheets of coarse-grained rock whose composition corresponds to liquid formed by approximately one-third fractional crystallization of the host basalt. Several mechanisms have been proposed for separating this liquid from the basalt, one of which involves expulsion of liquid from a crystal mush that undergoes compaction in the lower part of the flow⁷. Compaction can occur only after crystals interconnect to form a network but before the network grows too strong to be deformed. Previous partial melting experiments¹² on the thick Jurassic Holyoke flood basalt of Connecticut and Massachusetts indicate that such a network forms when this basalt is only 25% crystallized.

Here we present the results of experiments that indicate that the crystal mush in the Holyoke basalt owed its strength to the presence of a remarkable three-dimensional (3D) network of plagioclase crystal chains. We have determined the shape of this network from 3D models constructed from serial sections through partially melted samples of the basalt. The deformation of the network that resulted from compaction during solidification of the lava can be measured from the chain distribution in the serial sections. These measurements provide a direct means of determining the amount of compaction that has occurred during the solidification of an igneous body.

Partial melting experiments

Samples of the Holyoke basalt have been partially melted so that the textural relations between the refractory minerals involved in compaction could be studied. By partly melting and then rapidly quenching the samples, the low-melting, late-crystallizing minerals are converted to glass in which the refractory minerals are

embedded. One-centimetre cubes of the rock were heated in graphite crucibles under a reducing atmosphere. Graphite is not wetted by silicate melts and therefore does not wick away the liquid from the partially melted rock. The furnace atmosphere was not controlled precisely, but the cores of the rock cubes remained close to their intrinsic oxygen fugacity (quartz–fayalite–magnetite buffer) for the duration of the experiments (<24 hours).

Previous melting experiments¹² reveal that cubes of basalt from the central (slowly cooled) part of this flow retain their cubic shape when at least 70% melted, which indicates that a continuous network of crystals still exists at this high degree of melting. At 75% melt, however, the network is disrupted and the cubes collapse. The experiments also reveal that the network is permeated by interconnected channels through which the interstitial liquid travels easily, as indicated by an intrinsic permeability of $2 \times 10^{-10} \text{ m}^2$ at 34% crystallinity.

New experiments show that this behaviour characterizes the entire flow apart from the basalt near the margins. This difference is illustrated in an experiment in which three samples, the first from 1 m, the second from 47 m, and the third from 74 m above the base of the 174-m-thick flow at Tariffville, Connecticut, were heated simultaneously at 1,112 °C for 14 hours. All are fine-grained and consist of plagioclase laths and equant grains of augite and pigeonite with interstitial magnetite and granophyre. They have similar, but not identical, compositions (Table 1). The first is close to the initial magma composition calculated from mass balance through the flow. The second is interpreted to have formed from crystal mush that underwent 28% compaction, as indicated by its depletion in elements that are concentrated in the melt (for example, Ti) and enrichment in elements that fractionate into early crystallizing

Table 1 Composition (wt %) of Holyoke basalt

Height	1 m	47 m	74 m	Initial magma
SiO ₂	53.30	52.65	51.71	53.25
TiO ₂	1.05	0.80	1.10	0.95
Al ₂ O ₃	14.10	14.53	13.60	14.58
Fe ₂ O ₃	2.28	1.77	2.47	2.28
FeO	10.43	9.55	11.33	9.65
MnO	0.21	0.18	0.22	0.20
MgO	5.49	6.63	5.29	5.86
CaO	9.04	9.55	10.68	9.54
Na ₂ O	2.76	2.73	2.64	2.79
K ₂ O	1.13	0.67	0.71	0.72
P ₂ O ₅	0.12	0.09	0.14	0.05
Total	99.91	99.15	99.89	99.86

Holyoke basalt was analysed by X-ray fluorescence analysis at the given heights above base of flow, and initial magma was calculated from mass balance through 174-m-thick flow.

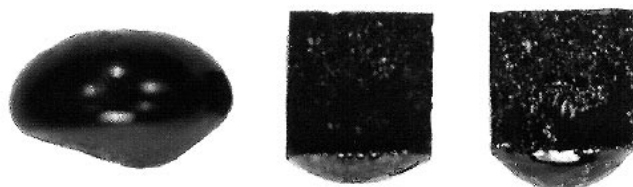


Figure 1 Three 1-cm cubes of basalt from 1, 47, and 74 m above the base of the Holyoke basalt after heating at 1,112 °C. Cubes were placed over circular holes drilled through the base of graphite crucibles. The 1-m sample melted, drained through the hole and formed a blob in an underlying crucible. The other two retained their cubic shape because of the presence of plagioclase crystal chains.

minerals (such as Mg). The third is interpreted to have formed from mush that was enriched with residual liquid (diluted crystal mush). Each cube was placed over a 1-cm-diameter hole drilled in the base of the crucible.

After heating, modal analyses indicated that the three samples contained 69, 63 and 73% melt, respectively. Despite the similar degrees of melting, the cubes behaved differently (Fig. 1): the first lost all strength and flowed through the hole in the base of the crucible to form a blob in an underlying crucible; the other two, however, retained their sharp cubic shape despite liquid draining from their interior to form lenses on their lower surface. This difference in behaviour is determined by the mineral distributions in these rocks. In the 1-m sample, plagioclase and pyroxene crystals and patches of late-crystallizing magnetite and granophyre are evenly distributed. Consequently, when the rock is partly melted, liquid forms between most grains, and the rock loses its strength even at low degrees of melting. In the other two samples, the refractory plagioclase crystals are clustered together in monomineralic chains that form a 3D network (Fig. 2). Plagioclase crystals in the partially melted rock are invariably separated from pyroxene crystals by zones of melt. If a pyroxene crystal were present in a plagioclase chain, melt would form between it and the plagioclase, and the chain would be disrupted. This never occurs: the chains contain only plagioclase crystals. It should be noted that the chains are not produced in the melting experiments. Chains are visible in the original unmelted basalt but are easier to recognize in the partially melted rock. All samples of the basalt from more than a couple of metres from the contact of the flow contain plagioclase chains.

Because of their persistence to high temperatures in the melting experiments, we infer that the plagioclase chains have crystallized early. It is possible, however, that the melting experiments do not reverse the natural crystallization sequence, and the persistence of plagioclase chains in the experiments is purely kinetic. Electron-microprobe analyses have been done along the length of chains to determine at what composition, and thus stage of crystallization, the plagioclase crystals linked together to form chains. One such traverse is shown in Fig. 3, where 14 crystal contacts are intersected. Two problems arise in interpreting this data—not all crystals need have attached at the same time (same composition), and sectioning presents a problem with the third dimension (the section may not intersect the initial point of contact). Nevertheless, the following general conclusions can be drawn. The phenocrysts have rounded cores of $\sim\text{An}_{55}$ surrounded by rims that zone outward to An_{65} , which is the core composition of the groundmass plagioclase crystals that make up the chains. These groundmass crystals are

normally zoned, with the highest anorthite content at the boundaries between crystals in the chains being $\sim\text{An}_{59}$. Thus, the chains are probably made from early crystallizing plagioclase crystals that remained freely suspended until their composition reached An_{59} .

3D models of plagioclase-chain network

The continuity of the chain-forming network is demonstrated by how well the third cube in Fig. 1 retained its shape, despite being 73% melted. Imaging this 3D network in 2D sections cut through a partially melted cube, however, is difficult. The chains in Fig. 2 are clearly visible because they parallel the plane of sectioning. In general, many chains will pierce this plane and will not be as easily recognized. Serial sections of the cubes must be used to obtain a 3D image of the network. This is time-consuming work, and so far, 3D images have been obtained for only two samples, but they do represent extreme rock types. They are from 47 and 74 m above the base of the flow at Tariffville and are interpreted to have formed from the most compacted and most diluted crystal mushes, respectively. One centimetre cubes of these rocks were heated until they were $\sim 70\%$ melted. They were then serially thinned by grinding away 0.1–0.2-mm-thick layers. Each new surface was polished and then etched with HF fumes, which helps distinguish between glass (blue), plagioclase (brown), and pyroxene (white) in reflected light (Fig. 2). A composite image of each layer was created by computer from 35 separate video frames obtained using a $3\times$ objective lens on a reflecting light microscope. The phase distributions were then mapped in each section, and from this a 3D model of the crystal mush was constructed. Figure 4 shows the distribution of plagioclase crystals in the serial sections from the two samples. This distribution becomes partly obscured when the sections are stacked on top of each other in a 2D diagram; they are much clearer when viewed sequentially in rapid succession (for details, see Fig. 4 legend).

The most striking feature seen in the serial sections (Fig. 4) is that the refractory plagioclase crystals are all part of monomineralic chains, whereas the pyroxene crystals are separate individuals between the chains. Although chains may include plagioclase phenocrysts, they are composed predominantly of groundmass crystals, which are less than 0.5 mm long and 0.1 mm wide in both the 47- and 74-m samples. The chains are several crystals wide and average 12 crystals per millimetre of chain length. They branch every 1 or 2 mm to create the 3D network. At branching points, the chains typically thicken.

Although the networks in the diluted and compacted samples (Fig. 4a and b, respectively) are similar, there are significant differences. Chains in the diluted sample form a more open network, that is, the melt channels are larger than in the compacted sample.

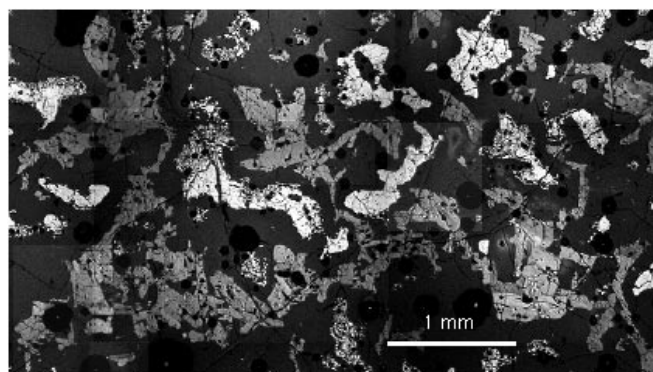


Figure 2 Continuous network of plagioclase chains in partially melted basalt from 47 m above base of flow at Tariffville. In this polished section, which was etched with hydrogen fluoride fumes, pyroxene is white, plagioclase light grey, and glass dark. Plagioclase crystals form a continuous network of chains that can be traced from the upper and lower left to the lower right side of the field.

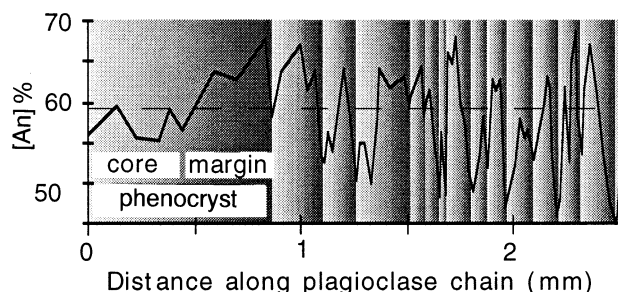


Figure 3 Anorthite content of plagioclase crystals in traverse along length of chain. Individual crystals are shown by shading. The maximum anorthite (An) content in the margin of the phenocryst and in the core of the groundmass crystals is $\sim\text{An}_{65}$. The horizontal dashed line at An_{59} represents the maximum An content where crystals come in contact.

The chains in the dilated sample are also randomly oriented, whereas those in the compacted sample are weakly aligned horizontally, which results in the channels between the chains being distinctly elongate in the horizontal plane. In addition, the chain segments in the compacted sample tend to be longer and straighter in the horizontal direction than in the vertical direction.

Deformation of the plagioclase-chain network

Because the network forms at such an early stage of crystallization (25%), it provides an ideal strain marker to track compaction of the crystal mush. The amount of compaction should be measurable from the anisotropy of the network, but the network is relatively irregular and not easily parameterized. However, if a mush had not undergone compaction, traverses in vertical and horizontal directions would be expected to intersect the same number of chains. In the sample shown in Fig. 4a, this is in fact the case, with 5.2 chains per cm being intersected in both directions. If a crystal mush has undergone compaction, however, the frequency at which chains are intersected in vertical and horizontal traverses would be expected to differ. In the sample shown in Fig. 4b, the frequency along horizontal traverses averages 5.15 chains per cm, not unlike the values in Fig. 4a, but in the vertical direction, the average is 6.25 chains per cm. These differences are consistent with the weak horizontal chain alignment seen in Fig. 4b and for the contrasting shape of the channels between the chains seen in Fig. 4a,b. If the horizontal frequency in Fig. 4b is taken as the pre-compaction value in both the vertical and horizontal directions, a frequency of 6.25 chains per cm in the vertical direction indicates 18% compaction.

During compaction, the interstitial liquid would have been displaced upwards through the mush and the content of incompatible elements in the compacted rock decreased. The composition of the 47-m sample was previously interpreted⁷ to indicate 28% compaction, that is, 10% more than is accounted for by the deformation of the network. By measuring the deformation of the plagioclase network, it is possible to draw the important conclusion that some process in addition to compaction must have been involved in the differentiation of this rock. At the moment this mechanism is unknown, but could involve enrichment in phenocrysts before network formation, infiltration metasomatism, or thermal and compositionally driven convection through the highly permeable mush.

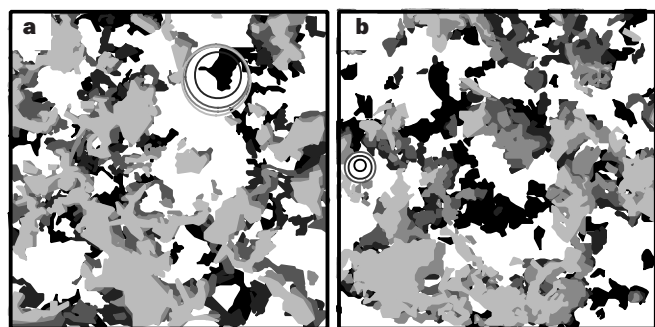


Figure 4 Serial sections showing monomineralic plagioclase chains in partially melted basalt. Serial sections are from **a**, 76 m, and **b**, 47 m above flow base. Both fields of view are 5 mm wide and 1 mm deep, with the shade of grey increasing with depth. Circles are outlines of bubbles in glass. The 74-m sample is interpreted to have been formed from dilated crystal mush and the 47-m sample from compacted mush. The serial sections can be viewed sequentially in animated mode on the World-Wide Web at <http://fermat.geol.uconn.edu/~philpott/>.

Theoretically, compaction of a crystal mush can cause a weak alignment of plagioclase crystals^{13,14}. However, in the most compacted sample from the Tariffville section (Fig. 4b), no such alignment is found. This lack of alignment is probably due to the fact that the crystals are bound together into chains and thus were not free to rotate independently of the chains. However, because chains exhibit a weak horizontal alignment in the compacted sample, a preferred orientation of plagioclase crystals might be expected if the crystals have a preferred orientation relative to the chain they constitute. To test this, the orientation of the (010) plane of individual crystals was measured relative to the axis of the local chain. Measurements were made on a segment of the network shown in Fig. 2, which contains 313 continuously linked crystals. Of these crystals, 100 had orientations that fell outside the angular range accessible to the universal stage that was used to make the measurements. The frequency of angles exhibited by the measurable crystals is shown in the histogram in Fig. 5: the solid line indicates the distribution expected from randomly oriented crystals (taking into account the blind region caused by the restricted tilt angle of the universal stage). The plagioclase crystals in the chain show no significantly preferred orientation with respect to how they are attached to the chain.

Formation of plagioclase chains

Plagioclase chains are not found in the chilled margins of the Holyoke basalt, so they are not the product of rapid quenching. Nor are they the product of dendritic crystal growth, because electron microprobe analyses indicate that the crystals constituting a chain originally existed as separate crystals while they grew from An₆₅ to An₃₉. Plagioclase crystals must therefore be attracted to one another once they have formed. The tendency for crystals to cluster together is a well known phenomenon for which Vogt¹⁵ coined the term 'synneusis' (swimming together). The surface energy on plagioclase crystals is the most likely cause for this clustering. The role of surface energies in determining the textural relations between crystals and melt when only small quantities of melt are present is well recognized^{16–18}. When plagioclase crystals are completely surrounded by melt, their surface free energy can be lowered by crystals becoming attached to one another. The fact that plagioclase crystals link together to form chains in the Holyoke basalt and pyroxene crystals do not, indicates that the attractive

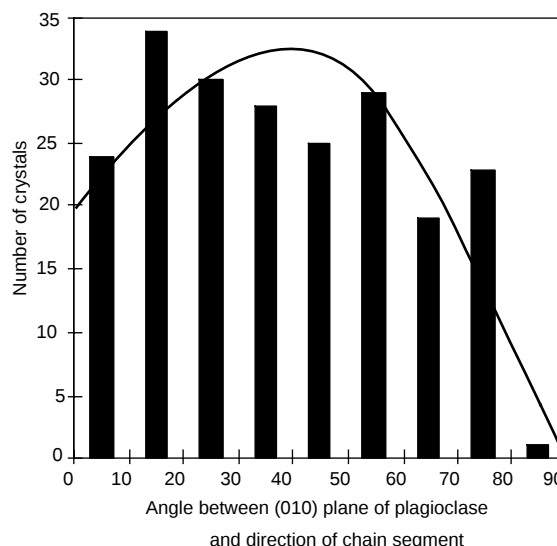


Figure 5 Universal stage measurements of the angular distribution of the (010) plane of 213 plagioclase crystals in a chain relative to the chain axis. The solid line indicates the expected distribution of randomly oriented crystals, with the drop-off at low angles resulting from the tilt restriction imposed by the universal stage.

force is selective. Plagioclase crystals in a basaltic melt would be expected to have higher surface energies than pyroxene crystals because of their greater structural mismatch with the liquid. Clustering of plagioclase crystals would therefore bring about greater reduction of surface energy than clustering of pyroxene crystals.

Plagioclase chains in plutonic rocks

The plagioclase chains in the Holyoke basalt would not have been discovered if the mush in the partial melting experiments had not been examined to find what gave it strength at high degrees of melting. Once found, however, the chains were identifiable in the unmelted rock, but they are far less prominent, because of the presence of later crystallizing minerals, including more sodic plagioclase. The fine grain size of the Holyoke basalt also increases the likelihood that parts of the network of chains will be intersected in a thin section. If chains exist in coarser basaltic rocks, they will be more difficult to identify.

A cursory examination of the University of Connecticut rock collection indicates that plagioclase chains are found not only in slowly cooled basaltic flows but are common in many intrusive gabbroic rocks, including the coarsest of all, the Precambrian massif type anorthosites. For example, the Palisades Sill in New Jersey, which from chemical evidence underwent compaction¹⁹, is similar in age and composition to the Holyoke basalt and contains similar but coarser plagioclase chains. Karoo dolerites in South Africa also contain these chains (as serendipitously illustrated on the cover of ref. 20). Much coarser, but similar-looking chains occur in the norite from the Stillwater Complex, Montana, and the plagioclase clots described from the olivine gabbro in this intrusion²¹ may be a related phenomenon. At a still coarser grain size, chains of plagioclase crystals are present in the Proterozoic noritic anorthosite from Lake St John, Quebec; these are beautifully displayed in the so-called 'black granite' from the quarry at St Gédion. This is one of the most widely used facing stones in North America, and polished slabs of this rock are available for inspection on many buildings. The chains in this rock, like those in the Holyoke basalt, are several crystals wide, but the individual crystals are up to a centimetre wide and several centimetres long. The chains also branch to form a 3D network, but at a much coarser scale. One difference is that the pyroxene crystals between the chains in the anorthosite are optically intergrown with smaller plagioclase crystals, whereas they form individual crystals in the Holyoke basalt. The plagioclase crystals in these optically patches also exhibit a strong horizontal alignment, whereas the much coarser plagioclase crystals in the chains, like those in the Holyoke basalt, are at oblique angles because the chains tend to be horizontal.

Implications

Formation of a crystalline network is one of the most important steps in the solidification of magma. Near the margins of igneous bodies, solidification fronts are attached to the walls through a crystal network. The strength of this network, which depends in large part on the degree of crystallization, affects the stability of roofward solidification fronts and the range over which compaction can occur in floorward solidification fronts²².

The serial sections through the partially melted samples of Holyoke basalt demonstrate that a network of chains of plagioclase crystals was present in the slowly cooled parts of this flow after only 25% crystallization. At such low crystallinity, the network was weak and highly permeable. Consequently, even though the density contrast between the bulk solids and liquid was small, the crystal mush was able to undergo significant compaction in the lower third of the flow with expulsion of the liquid upward into the central part of the flow. As the crystallinity increased above 35%, however, the network became too strong to be deformed and compaction ceased¹². The change in the strength of the crystal network therefore

creates a crystallization interval during which compaction can occur. In the Holyoke basalt, this interval is between 25 and 35%, but in basalts of different composition it may shift, depending on when the network develops. Nonetheless, the fact that compaction occurs during a definite interval of crystallization—a compaction window—means that differentiation by this mechanism will be stepwise rather than continuous. Most segregation sheets in thick flood-basalt flows, for example, have compositions that indicate they separated from the host basalt after ~35% crystallization of the host.

Although the network of plagioclase chains has been documented carefully only in the Holyoke flood basalt, the presence of similar, but coarser, networks in plutonic rocks suggests that plagioclase crystal networks are a ubiquitous feature of slowly cooled basaltic magmas. The exact stage of crystallization at which these networks form undoubtedly depends on the composition of the magma and its liquidus phase relations. Plagioclase must be a liquidus phase, and in most basaltic magmas this occurs at an early stage of crystallization. Based on the evidence from the Holyoke basalt, which is a normal quartz tholeiite (Table 1), plagioclase crystal-chain networks probably form early enough in many basaltic magmas to play a significant role in their differentiation.

If a network of plagioclase crystal chains does form early in the crystallization of a basaltic magma, the evidence from the Holyoke basalt indicates that the network provides an ideal strain marker with which to track the compaction of the crystal mush. The fact that this allows the degree of compaction to be determined directly and independently of the chemistry of the rock means that the chemical data could be used to extract information about other processes that may be active during compaction of basaltic crystal mush. □

Received 16 January; accepted 25 June 1998.

- Cornwall, H. R. Differentiation in lavas of the Keweenaw series and the origin of the copper deposits of Michigan. *Geol. Soc. Am. Bull.* **62**, 159–202 (1951).
- Lindsley, D. H., Smith, D. & Haggerty, S. E. Petrography and mineral chemistry of a differentiated flow of Picture Gorge Basalt near Spray, Oregon. *Carnegie Inst. Wash. Yearb.* **69**, 264–285 (1971).
- Dostal, J. & Greenough, J. D. in *Eastern North American Mesozoic Magmatism* (eds Puffer, J. H. & Ragland, P. C.) 149–160 (Spec. Pap. 268, Geol. Soc. Am., 1992).
- Greenough, J. D. & Dostal, J. Layered rhyolite bands in a thick North Mountain basalt flow: The products of silicate liquid immiscibility? *Mineral. Mag.* **56**, 309–318 (1992).
- Puffer, J. H. & Horter, D. L. Origin of pegmatitic segregation veins within flood basalts. *Geol. Soc. Am. Bull.* **105**, 738–748 (1993).
- Wheelock, M. M. & Marsh, B. D. Silicic segregations in basaltic igneous bodies: Their nature, occurrence and conditions of formation. *EOS Abstr. Spring 1994* **75**, No. 16, 365 (1994).
- Philpotts, A. R., Carroll, M. & Hill, J. M. Crystal-mush compaction and the origin of pegmatitic segregation sheets in a thick flood-basalt flow in the Mesozoic Hartford basin, Connecticut. *J. Petrol.* **37**, 811–836 (1996).
- Richter, D. H. & Moore, J. G. *Petrology of the Kilauea Iki Lava Lake, Hawaii* (Prof. Paper 537-B, US Geol. Surv., 1966).
- Moore, J. G. & Evans, B. W. The role of olivine in the crystallization of the prehistoric Makaopuhi tholeiitic lava lake, Hawaii. *Contrib. Mineral. Petrol.* **15**, 202–223 (1967).
- Wright, T. L. & Okamura, R. T. *Cooling and Crystallization of Tholeiitic Basalt, 1965 Makaopuhi Lava Lake, Hawaii* (Prof. Paper 1004, US Geol. Surv., 1977).
- Helz, R. Crystallization history of Kilauea Iki lava lake as seen in drill core recovered in 1967–1979. *Bull. Volcanol.* **43**, 675–701 (1980).
- Philpotts, A. R. & Carroll, M. Physical properties of partly melted tholeiitic basalt. *Geology* **24**, 1029–1032 (1996).
- Higgins, M. The origin of laminated and massive anorthosite, Sept Iles layered intrusion, Quebec, Canada. *Contrib. Mineral. Petrol.* **106**, 340–354 (1991).
- Nicolas, A. Kinematics in magmatic rocks with special reference to gabbros. *J. Petrol.* **33**, 891–915 (1992).
- Vogt, J. H. L. The physical chemistry of the crystallization and magmatic differentiation of igneous rocks. *J. Geol.* **29**, 318–350 (1921).
- Waff, H. S. & Bulau, J. R. Equilibrium fluid distributions in an ultramafic partial melt under hydrostatic stress conditions. *J. Geophys. Res.* **84**, 6109–6114 (1979).
- Cooper, R. F. & Kohlstedt, D. L. Interfacial energies in the olivine–basalt system. *Adv. Earth Plan. Sci.* **12**, 217–228 (1982).
- Hunter, R. H. in *Origins of Igneous Layering* (ed. Parsons, I.) 473–503 (Reidel, Dordrecht, 1987).
- Shirley, D. N. Differentiation and compaction in the Palisades Sill, New Jersey. *J. Petrol.* **28**, 835–865 (1987).
- Philpotts, A. R. *Principles of Igneous and Metamorphic Petrology* (Prentice Hall, NJ, 1990).
- Meurer, W. P. & Boudreau, A. E. Petrology and mineral compositions of the middle banded series of the Stillwater Complex, Montana. *J. Petrol.* **37**, 583–607 (1996).
- Marsh, B. D. Solidification fronts and magmatic evolution. *Mineral. Mag.* **60**, 5–40 (1996).

Acknowledgements. We thank B. D. Marsh for useful comments. This work was supported by a grant from NSF.

Correspondence and requests for materials should be addressed to A.R.P. (e-mail: philpotts@geol.uconn.edu).

Crystal structure of a SNARE complex involved in synaptic exocytosis at 2.4 Å resolution

R. Bryan Sutton*, Dirk Fasshauer†, Reinhard Jahn† & Axel T. Brunger*

* The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA

† Department of Neurobiology, Max-Planck-Institute for Biophysical Chemistry, D-37077 Göttingen, Germany

The evolutionarily conserved SNARE proteins and their complexes are involved in the fusion of vesicles with their target membranes; however, the overall organization and structural details of these complexes are unknown. Here we report the X-ray crystal structure at 2.4 Å resolution of a core synaptic fusion complex containing syntaxin-1A, synaptobrevin-II and SNAP-25B. The structure reveals a highly twisted and parallel four-helix bundle that differs from the bundles described for the haemagglutinin and HIV/SIV gp41 membrane-fusion proteins. Conserved leucine-zipper-like layers are found at the centre of the synaptic fusion complex. Embedded within these leucine-zipper layers is an ionic layer consisting of an arginine and three glutamine residues contributed from each of the four α -helices. These residues are highly conserved across the entire SNARE family. The regions flanking the leucine-zipper-like layers contain a hydrophobic core similar to that of more general four-helix-bundle proteins. The surface of the synaptic fusion complex is highly grooved and possesses distinct hydrophilic, hydrophobic and charged regions. These characteristics may be important for membrane fusion and for the binding of regulatory factors affecting neurotransmission.

Fusion of a vesicle with its target membrane is mediated by a set of conserved proteins collectively referred to as SNAREs (soluble NSF-attachment protein (SNAP) receptors; NSF is *N*-ethylmaleimide-sensitive fusion protein)^{1,2}. In exocytosis of synaptic vesicles, these SNAREs include the plasma-membrane-associated proteins syntaxin and SNAP-25 (synaptosome-associated protein of relative molecular mass 25K), and the vesicular protein synaptobrevin. Although the function of the SNAREs in membrane fusion is not fully understood, it must be fundamental because proteolytic cleavage of the SNAREs by clostridial neurotoxins inhibits neurotransmission³. SNARE proteins spontaneously assemble into a stable ternary complex that has a melting temperature (T_m) of 90 °C^{4–6}. According to present theories, after the SNARE proteins form a complex, the vesicle and target membranes fuse, thereby releasing neurotransmitter into the synaptic cleft². After fusion, the SNARE complex can then recruit a member of the SNAP family^{7,8}. Addition of SNAPs to the SNARE complex allows binding of NSF, an ATPase that catalyses the dissociation of the ternary SNARE

complex, thereby priming the SNAREs for another round of fusion^{4,5,9,10}.

Because the assembly of synaptic SNAREs is important in mediating neurotransmission, information about the SNARE-complex structure is crucial to understanding the detailed interactions among SNAREs and with other regulatory factors. Here we report the 2.4 Å crystal structure of the synaptic fusion complex consisting of the cytoplasmic domain of synaptobrevin-II, the carboxy-terminal H3 domain of syntaxin-1A, and the amino- and carboxy-terminal domains of SNAP-25B (ref. 11). We have shown previously that this core complex resembles the full-length complex with respect to assembly, disassembly, and biophysical properties^{6,11}.

Structure determination

The core synaptic fusion complex was obtained by limited proteolysis, mass-spectrometry analysis, and bacterial expression of the fragments¹¹. The complex consists of syntaxin-1A residues Gly 180–Arg 262 (Sx), synaptobrevin-II residues Met 1–Met 96 (Sb), and the

Table 1 Crystallographic data, phasing and refinement

Crystal*	<i>a</i> (Å)	<i>b</i> (Å)	<i>c</i> (Å)	λ (Å)	<i>d</i> _{min} (Å)	Observed reflections (<i>n</i>)	Unique reflections (<i>n</i>)	Completeness (%)	$\langle I \rangle / \langle \sigma \rangle$	<i>R</i> _{sym} † (%)
Native	100.54	112.71	200.77	1.0332	2.4	309,408	42,980	99.1 (98.8)	17.4	7.9 (32.9)
Iodine	100.44	112.77	199.00	1.0332	3.0	81,645	42,094	96.2 (89.1)	12.0	6.7 (46.2)
SeMetSn1	98.72	111.07	198.84	0.9797	3.0	160,622	46,665	99.7 (99.4)	17.0	8.1 (37.7)
SeMetSx	100.11	113.13	198.93							
λ ₁ (low-energy remote)				1.0688	3.2	120,963	25,261	97.9 (97.9)	16.9	6.4 (26.9)
λ ₂ (inflection point)				0.9801	3.2	110,188	35,008	97.3 (95.9)	13.0	7.9 (32.0)
λ ₃ (anomalous peak)				0.9797	3.2	112,147	35,072	97.3 (96.1)	13.9	7.3 (32.0)
SeMetSn1Sn2	99.88	112.95	198.64							
λ ₁ (low-energy remote)				1.0688	3.0	137,029	42,024	96.5 (86.3)	16.7	6.0 (28.3)
λ ₂ (inflection point)				0.9800	3.0	129,455	41,257	95.1 (82.2)	13.8	7.2 (38.5)
λ ₃ (anomalous peak)				0.9795	3.0	129,052	41,449	95.0 (81.6)	12.6	7.5 (42.0)
SeMetSn1Sn2	99.94	112.79	198.61							
λ ₁ (low-energy remote)				1.0688	3.3	115,718	31,945	97.8 (98.5)	14.7	8.1 (28.3)
λ ₂ (inflection point)				0.9800	3.3	113,142	31,613	97.3 (97.9)	13.8	10.4 (35.3)
λ ₃ (anomalous peak)				0.9795	3.3	111,867	31,456	97.4 (97.7)	11.3	10.8 (38.0)

* Sn1Sn2, synaptic fusion complex with selenomethionine (SeMet) labels on both Sn1 and Sn2; Sn1SnSx, synaptic fusion complex with SeMet labels on Sn1, Sn2 and Sx. Values in parentheses are for the highest resolution bins.

† $R_{\text{sym}} = \sum_i \sum_h |I_i(h) - \langle I(h) \rangle| / \sum_i \sum_h I_i(h)$, where $I_i(h)$ is the *i*th measurement and $\langle I(h) \rangle$ is the weighted mean of all measurements of $I(h)$ for Miller indices *h*.

N- and C-terminal fragments of SNAP-25B, residues Met 1–Lys 83 (Sn1) and residues Val 120–Gly 206 (Sn2), respectively. We obtained crystals that exhibited limited diffraction, variable crystal quality, and high mosaicity. Cryotechniques combined with crystal annealing against increasing methylpentanediol (MPD) concentrations were essential in obtaining useful diffraction to a minimum Bragg spacing of 2.4 Å (Table 1). The crystals are of space group *I*222 ($a = 100.54$ Å, $b = 112.71$ Å, $c = 200.77$ Å; Table 1), with three synaptic fusion complexes and 54.8% solvent in the crystallographic asymmetric unit.

The complex was solved by a generalization of the method of multiwavelength anomalous dispersion (MAD) phasing¹². Because of the size of the asymmetric unit and the limited crystal quality, multiple MAD experiments were required, each using a different combination of native and selenomethionine-labelled SNARE complexes (Table 1). Combination of the phases of the individual MAD experiments resulted in a significantly improved experimental electron-density map, because the directions of the heavy-atom structure factors are different for each combination of selenomethionine labels (Fig. 1 and Table 1 in Supplementary Information). Furthermore, the distribution of syntaxin-1A selenomethionine labels in one of the data sets allowed determination of the labelled methionine positions using an automated Patterson search method. The remaining sites were found by difference Fourier techniques. This 'multi-MAD' technique could be generally useful for structure solution of large macromolecular complexes with limited crystal quality and variability.

High-quality experimental phases allowed unbiased comparison of the three independent synaptic fusion complexes in the asymmetric unit. Of the three complexes, two are related by an approx-

imate two-fold operation and the third complex is related by improper symmetry. These associations are mediated by ordered strontium ions and intercomplex sidechain interactions. Two strontium ions bridge the C termini of the synaptobrevin components of two antiparallel complexes. Another strontium ion bridges the syntaxin components of the two parallel orientated complexes. Although the synaptic fusion complex exhibits monomer–trimer equilibrium in solution¹¹, it is unknown whether these cation-mediated associations are related to any physiological function.

Overall topology

The synaptic fusion complex is arranged as a cylinder, 120 Å in length with circular cross-section (Fig. 2a). All four components of the heterotrimer are arranged in parallel, with the N termini at one end of the bundle and the C termini at the membrane-anchor end. The extreme N terminus of the core complex consists of a two-helix interaction between syntaxin and the Sn1 α -helix of SNAP-25. The central portion consists of a four-helix bundle with a left-handed superhelical pitch in which the synaptobrevin α -helix juxtaposes the syntaxin and Sn2 α -helices. The complex ends with a two-helix interaction between synaptobrevin and syntaxin at the C terminus.

The parallel orientation of synaptobrevin and syntaxin in the synaptic fusion complex has been shown by electron microscopy analysis¹³. However, parallel orientation of the two SNAP-25B α -helices was unexpected. In the full-length primary structure of SNAP-25B, a linker of 37 residues connects the Sn1 and Sn2 fragments. In the crystal structure, the N-terminal end of the Sn2 α -helix begins with a turn at residue Val 37, leaving 54 residues available to form a loop between the two SNAP-25 α -helices. This loop is probably palmitoylated *in vivo* at conserved cysteine residues

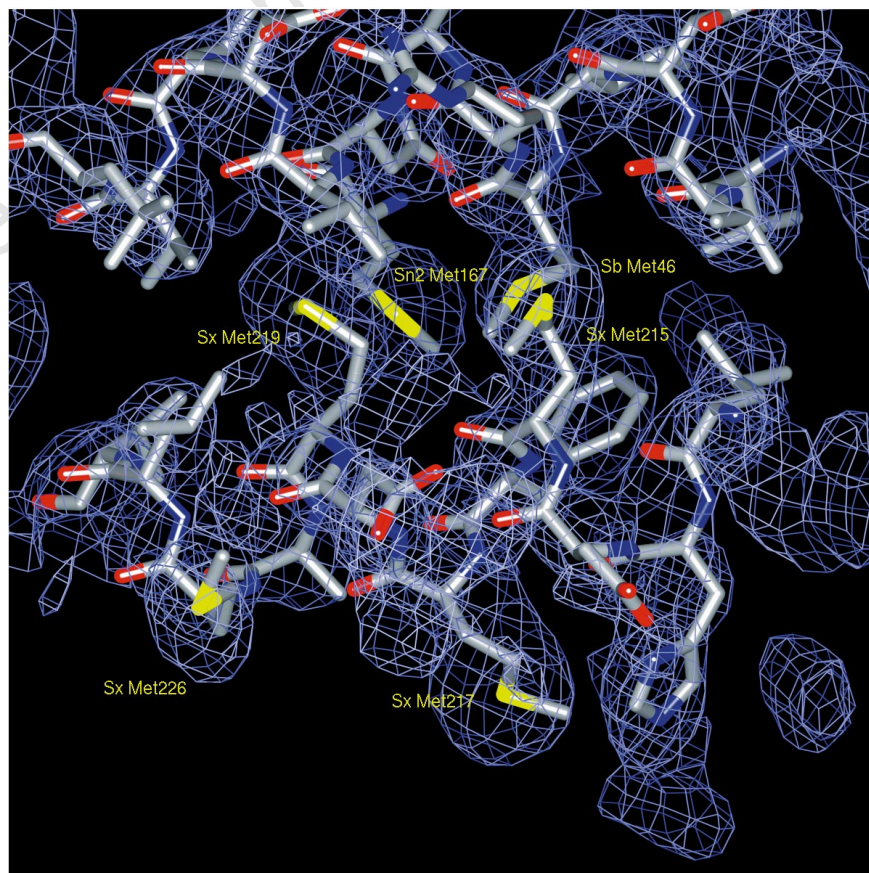


Figure 1 Electron-density map of the synaptic fusion complex obtained by multi-MAD phasing, showing a methionine cluster. Multi-MAD phases were obtained from phase combination of the Sx, Sn1Sn2 and Sn1Sn2Sx MAD data sets at 3 Å

resolution, followed by density modification and phase extension to 2.4 Å resolution. The map was contoured at 1σ . The structure was superimposed as liquorice bonds.

and used as a membrane anchor¹⁴. As it is easily cleaved by proteases, the loop between SNAP-25 α -helices may be unstructured. We were concerned that, in the proteolytically cleaved synaptic complex, the SNAP-25B fragments might orientate themselves in the SNARE complex independently of each other, as the fragments are unrestrained. The parallel orientation of the Sn1 and Sn2 α -helices was shown by the location of the selenomethionine labels and by fitting the Sn1 and Sn2 structures to the experimental electron-density map. In addition, all of the Sn1 and Sn2 α -helices observed in the asymmetric unit possess the same parallel orientation. If there were a mixture of orientations in solution, one might expect the asymmetric unit in the crystal to reflect that distribution. Modelling of the entire SNAP-25B protein confirmed that the 54-residue loop between the two SNAP-25 α -helices was sufficient to span the 84 Å connecting the two fragments, thereby allowing a parallel orientation of α -helices.

Four-helix-bundle structure

Each of the three synaptic fusion complexes in the asymmetric unit possesses a different superhelical bend (Fig. 2b). We observed significant flexibility for both backbone and side-chain atoms in each of the molecules in the asymmetric unit (Fig. 2b). The overall pitch of the four-helix bundle is not defined, as each of the components in the bundle possesses its own unique α -helical character (Fig. 2c). The radius of the four-helix bundle changes significantly over the length of the bundle (Fig. 2d). The variation in the radius is correlated with sidechain packing in the core of the

complex: the layers with the largest sidechain-packing volume show the largest radius.

Although the backbone topology of the complex (Fig. 2a) superficially resembles the tetrameric GCN4 four-helix-bundle structure¹⁵, several factors differentiate the synaptic fusion complex from the leucine-zipper or coiled-coil proteins described so far. First, the distribution of leucine, isoleucine and valine residues as core constituents is not uniform over the length of the synaptic fusion complex. The composition of many layers includes residues other than leucine, isoleucine and valine residues (Table 2 in Supplementary Information), contrasting with typical leucine zipper proteins¹⁶. Second, the geometry of most of the layers in the synaptic fusion complex deviates significantly from that seen in the centre of the GCN4 tetramer structure (Table 2 in Supplementary Information). The leucine-zipper geometry, as found in the tetrameric GCN4 structure, derives from a generalization of Crick's classical coiled-coil model¹⁷ to four-helix bundles. This model has planar and symmetrical layers whose normal is perpendicular to the superhelical bundle axis. Several regions in the individual proteins of the synaptic fusion complex are predicted to form a coiled-coil structure as they show heptad-repeat patterns (we used the Multi-Coil program¹⁸). These regions are from Leu60 to Trp90 of synaptobrevin-II, from Ala 5 to Arg 30 and from Leu 50 to Asp 80 of Sn1, and from Asn 169 to Gly 204 of Sn2. These patterns do not necessarily translate into a coiled-coil geometry for the entire four-helix bundle as the heptad-repeat patterns for the four individual fragments are not in register. As a consequence of these deviations

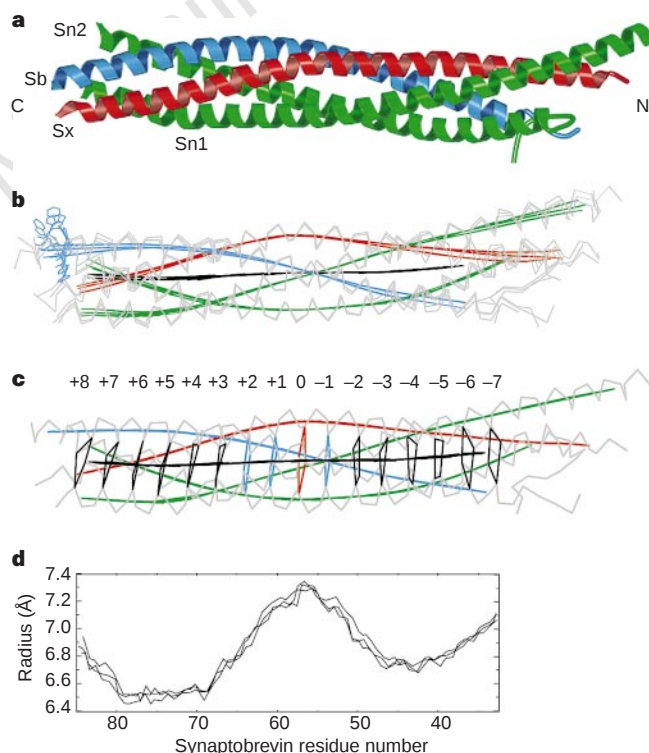


Figure 2 Topology and organization of the synaptic fusion complex. **a**, Backbone ribbon drawing of the synaptic fusion complex: blue, synaptobrevin-II; red, syntaxin-1A; green, SNAP-25B (Sn1 and Sn2). **b**, Conformational variability assessed by overlay of the three non-crystallographically related complexes. The backbones of layers -1 , 0 , $+1$ and $+2$ were superimposed with a pairwise r.m.s.d. of 0.3Å . The N and C termini of the synaptic fusion complex show significant variation among complexes (maximum r.m.s.d. around the mean position for backbone atoms is 1.7Å) and among side chains, such as Tyr88, Trp89, and Trp90 in synaptobrevin (blue). C α traces are in grey. **c**, Organization of

the synaptic fusion complex. C α traces (grey), local helical axes (blue, red and green for synaptobrevin-II, syntaxin-1A and SNAP-25B, respectively), the superhelical axis (black), and layers (0 , red; -1 , $+1$ and $+2$, blue; all others black; numbers refer to Table 2 in Supplementary Information) are shown for one of the three complexes in the asymmetric unit. Layers are indicated by virtual bonds between corresponding C α positions. **d**, Radii of the three synaptic fusion complexes in the asymmetric unit. The radius is defined as the average distance between the local helical axes and the central four-helix-bundle axis.

from standard coiled-coil heptad-repeat pattern and geometry, the α -helical curvature is unique for each of the four components (Fig. 2a). The angle at which each α -helix crosses the overall bundle axis, or the crossing angle, ranges from a minimum of about 5° in syntaxin and Sn1 to as much as 25° in synaptobrevin and Sn2. Crick¹⁷ predicted a crossing angle of 18° for dimeric coiled-coil proteins.

Residues from the four associating α -helices can be grouped into layers that are conserved in primary-structure alignments¹⁹ and which are sensitive to mutations^{20–23} (Fig. 2c). Layers '–1', '+1' and '+2' at the centre of the complex most closely follow ideal leucine-zipper geometry and amino-acid composition (Table 2 in Supplementary Information); each hydrophobic residue interacts in a plane. The layers are composed of leucine, isoleucine and valine residues and follow the packing characteristics of parallel, tetrameric leucine-zipper proteins¹⁵. However, in contrast to classical coiled-coil geometry, the tilt angles of the layers are between 8° and 14° with respect to the bundle axis (Fig. 2c, and Table 2 in Supplementary Information). Packing in the flanking regions of the central layers is typical of the more general class of α -helix-bundle proteins, as illustrated by the methionine cluster in Fig. 1.

Within the leucine-zipper layers, there is a highly conserved and completely buried '0' layer composed of Arg 56 from synaptobrevin-II, Gln 226 from syntaxin-1A, Gln 53 from Sn1 and Gln 174 from Sn2 (Fig. 3). The positively charged guanidino groups of the arginine residue interact with carboxyl groups from each of the three glutamine residues. The flanking leucine-zipper layers act as a water-tight seal to shield the ionic interactions from the surrounding solvent. This seal may further stabilize the four-helical oligomeric state and register of the complex by decreasing the local dielectric, thereby enhancing electrostatic interaction within the ionic layer. If another protein such as α -SNAP or NSF were able to puncture this seal, the ionic layer would be exposed to solvent, thereby increasing the local dielectric and weakening the interhelix interactions. This process may facilitate disassembly of the complex.

Surface structure

There are four shallow grooves on the surface of the synaptic fusion complex, formed by the association of the α -helices. In a manner

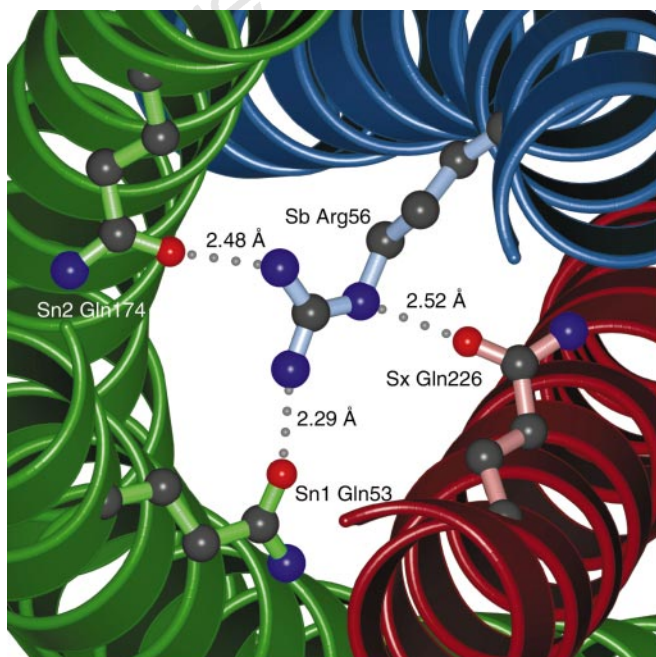


Figure 3 Ionic '0' layer of the synaptic fusion complex. Side chains involved in the layer are shown as balls and sticks; backbone is shown as a ribbon. The total buried surface area for the sidechain atoms in this layer is 742 Å².

similar to the interaction of transcription factors with DNA, effectors such as α -SNAP or complexin could use the grooves of the synaptic fusion complex as specific binding sites, possibly forming a higher-order helical bundle. Indeed, the N-terminal end of Sn2 fold back and interacts through hydrophobic contacts to one of the grooves of the four-helix bundle formed by the vesicle (v)-SNARE, synaptobrevin and the Sn2 α -helix of the target membrane (t)-SNARE, SNAP-25B (Fig. 4).

The arrangement of the t-SNARE components syntaxin, Sn1 and Sn2 in the four-helix bundle allows formation of a cradle into which the v-SNARE, synaptobrevin, may bind. In the synaptic fusion complex, the α -helices are knitted together by hydrophobic interactions in the core as well by several disperse hydrogen-bonding or salt-bridging sidechain interactions on the solvent-exposed surface. Analysis of these surface interactions shows that there are significantly fewer interactions originating from the v-SNARE than those originating from each of the t-SNAREs. There are only two side chains that form strong salt-bridge interactions from synaptobrevin to syntaxin. There are five such interactions between synaptobrevin and the neighbouring Sn2 α -helix. This indicates that most of the binding energy for synaptobrevin in the synaptic fusion complex comes from hydrophobic and ionic core packing interactions. The t-SNAREs, however, maintain 17 surface salt bridges between each other, 10 formed between the SNAP-25B components and 7 of which are between syntaxin and its adjoining SNAP-25B α -helix (Sn1). Most of these ionic interactions are between the 'e' and 'g' positions of the heptad-repeat scheme. These observations may explain why binary complexes containing syntaxin-1A and SNAP-25B are more stable than binary complexes containing synaptobrevin and either syntaxin-1A or SNAP-25B (ref. 11).

Neurotoxin cleavage

The SNARE proteins are the targets for the clostridial neurotoxins, including botulinum and tetanus neurotoxins³. All

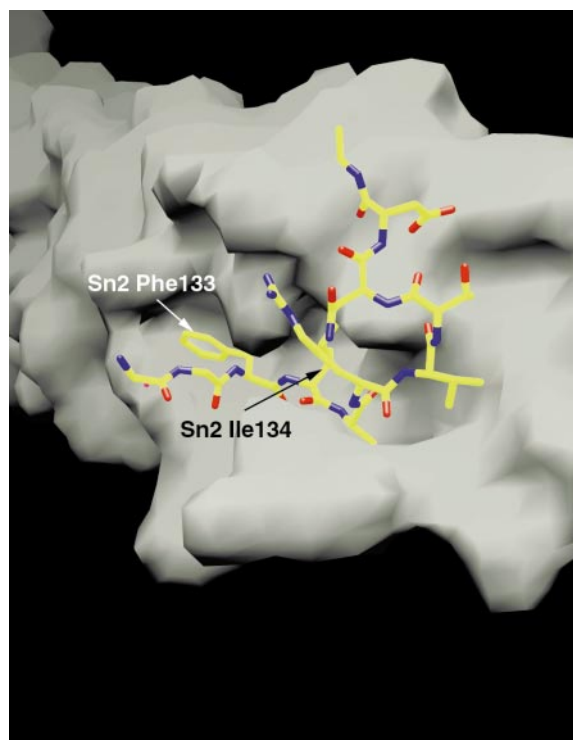


Figure 4 Close-up view of the loop involving Sn2 at the C-terminal end of the synaptic fusion complex. The SN2 loop continues as an extended peptide chain (shown as liquorice bonds) interacting with a hydrophobic groove in the synaptic fusion complex (shown as a molecular surface) formed by synaptobrevin and Sn2. Figure prepared with GRASP⁴⁸.

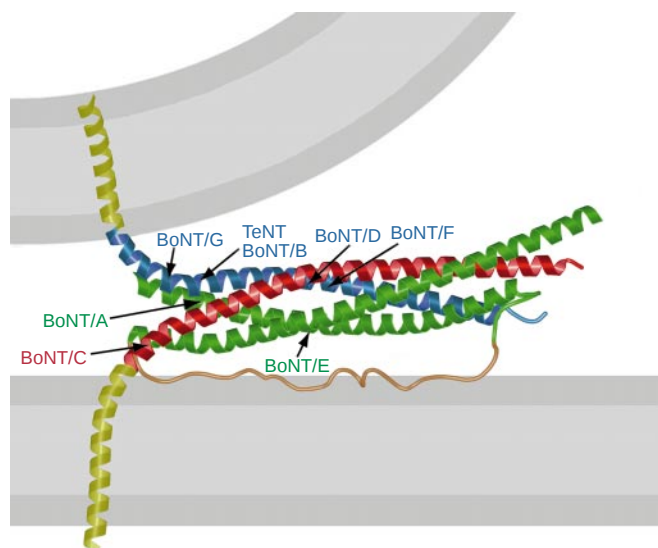


Figure 5 Hypothetical model of the synaptic fusion complex as it joins two membranes, and location of neurotoxin-mediated cleavage sites. We extended the synaptic fusion complex crystal structure to include the transmembrane domains (yellow) of syntaxin-1A (red) and synaptobrevin-II (blue), and the loop connecting the Sn1 and Sn2 fragments (green). The transmembrane domains and the linker to the Sx fragment are represented as α -helices. Hypothetical bends of the syntaxin and synaptobrevin α -helices were modelled close to the lipid bilayers. The loop between the Sn1 and Sn2 fragments was modelled as an unstructured polypeptide chain. The conformation of this loop is speculative. The loop between the Sn1 and Sn2 domains is shown in orange. We assumed that the lipid bilayers (grey) have a thickness of roughly 30 Å. The synaptobrevin-II neurotoxin-mediated cleavage site for tetanus toxin (TeNT) and botulinum toxin (BoNT) type B (BoNT/B) is between Gln 76 and Phe 77; for BoNT/F, between Gln 58 and Lys 59; for BoNT/G, between Ala 81 and Ala 82; and for BoNT/D, between Lys 59 and Leu 60. The syntaxin-1A BoNT/C cleavage site is between Lys 253 and Ala 254. Cleavage sites in SNAP-25B are between Asp 193 and Glu 194 for BoNT/E, and between Arg 176 and Gln 177 for BoNT/A.

neurotoxin-mediated cleavage sites are located between the C-terminal membrane anchors and the ionic '0' layer (Fig. 5). Only uncomplexed or partially assembled SNARE proteins can undergo proteolysis by neurotoxins²⁴. The fully assembled SNARE complex is resistant to proteolysis because either the protease-cleavage sites or the protease-recognition sites are protected upon complex formation. An intermediate, partially assembled state may exist *in vivo* before fusion; in this complex the α -helices at C-terminal end are separated (Fig. 5). As neurotoxins cleave synaptobrevin and syntaxin close to the membrane, cleavage may detach either the presynaptic membrane or the vesicle membrane from the prefusion complex. In contrast, the result of cleavage of Sn2 by the botulinum neurotoxins, type A or E, is less direct as the C terminus of Sn2 is not directly anchored to the membrane. Cleavage by these toxins may destabilize the four-helical bundle of the synaptic fusion complex in the C-terminal region²⁴, and may disrupt the ability of the complex to join membranes.

Mechanistic implications for fusion

The assembly of the synaptic fusion complex probably begins with the formation of the binary complex, consisting of SNAP-25 and syntaxin, on the target membrane. After binding of synaptobrevin to the binary complex, the membrane anchors of syntaxin and synaptobrevin are positioned on the same side of the complex (Fig. 5). Electrostatic calculations show a pronounced basic-charge distribution at the membrane-anchored end, or C terminus, of the synaptic fusion complex (Fig. 6). When the four α -helices coalesce

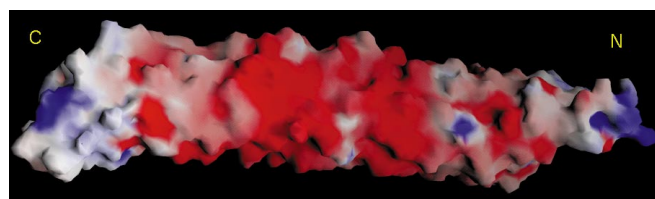


Figure 6 Surface plot showing the electrostatic potential of the synaptic fusion complex. Blue, positive charge; red, negative charge. Charges were obtained from the OPLS force field⁵⁰. The electrostatic surface was contoured between -10 kT/e and $+10$ kT/e. Figure prepared with GRASP⁴⁸.

into a complex immediately before fusion, the cumulative electrostatic potential may promote membrane fusion by affecting both membrane surfaces. Two tryptophan residues and a tyrosine are exposed to solvent at the C terminus of synaptobrevin-II in the crystal structure (Fig. 2b). These residues may associate with the membrane interface, as is often found in other membrane proteins²⁵, and may help in the fusion of vesicle and target membranes.

The inherent flexibility of the synaptic fusion complex may also contribute to its membrane-fusion activity. The predicted 20-residue transmembrane domain of synaptobrevin-II begins immediately at the C-terminal end of the Sb fragment. Likewise, the 23-residue transmembrane domain of syntaxin-1A and a 5-residue basic linker precede the Sx fragment. The transmembrane domains of syntaxin and synaptobrevin are probably α -helical, as the number of hydrophobic residues in the transmembrane domains would be sufficient to form a bilayer-spanning α -helix (Fig. 5). It is possible that the transmembrane α -helices would continue into the α -helices found in the synaptic fusion complex. If the synaptic fusion complex is involved in joining vesicles and target membranes, one would expect significant distortion or bending of the α -helices (Fig. 5). Indeed, flexibility of the α -helices is observed in both C-terminal and N-terminal regions (Fig. 2b). According to this model, the strain introduced by the α -helices may deform the lipid bilayers and promote lipid mixing.

The synaptic fusion complex is unlike the homo-oligomeric α -helical bundles that are found in the HIV/SIV gp41 (refs 26–28, 51) and the influenza virus haemagglutinin²⁹ crystal structures. These viral proteins may also be involved in promoting membrane fusion by a conformational change and subsequent antiparallel association of α -helices. In contrast, the synaptic fusion complex is a heterotrimeric, parallel four-helix bundle. The individual protein components of the synaptic fusion complex show little tendency to oligomerize, unlike the viral fusion proteins^{6,30}. This preference for heterotrimerization is determined by packing interactions in the core, flanking ionic side-chain interactions on the surface of the fusion complex, and the interactions in the ionic '0' layer. A common theme among recent models of fusion is emerging, namely the joining of two membrane surfaces through a proteinaceous agent. In the case of viral fusion, conformational changes in the fusion protein accomplish the juxtaposition of membranes³¹, whereas in the synaptic fusion complex, the assembly of the heterotrimer itself leads to fusion (Fig. 5). It remains to be determined whether the free energy released by the assembly of the synaptic fusion complex is sufficient to induce lipid mixing. Perhaps this process is assisted or regulated by accessory proteins, which could link Ca^{2+} -dependent exocytosis and the synaptic fusion complex. □

Methods

Protein production, crystallization and data collection. Characterization of the core synaptic fusion complex is described elsewhere¹¹. The four fusion-complex components were cloned into vector pET-28b (Novagen) and grown

in BL21(DE3) *Escherichia coli* cells (Novagen) with 50 mg ml⁻¹ kanamycin in a 10-l BioFlo3000 fermentor (New Brunswick Scientific) using enriched high-density media³² for Sb and defined media for the Sx, Sn1 and Sn2 components (T. J. Griffin, personal communication). Cells were induced with 0.8 mM isopropylthiogalactoside (IPTG) at an $A_{600} = 20.0\text{--}40.0$, collected by centrifugation, and flash-frozen in liquid nitrogen. Purification followed the previously published protocol¹¹. The iodine derivative was prepared by incubating Sb in 10 mM sodium iodide followed by addition of iodo-beads (Pierce) before complex formation. Selenomethionine-labelled proteins were obtained by expression in the B834λ(DE3) (Novagen) auxotroph, grown to $A_{600} = 2.0\text{--}3.0$ in selenomethionine-containing defined media³³, and purified in the presence of 10 mM dithiothreitol (DTT). Composition of the purified complex, iodine and selenomethionine incorporation were assessed by MALDI-TOF mass spectrometry (Perseptive).

Crystals of the complex were obtained by vapour diffusion at 29 °C using the sitting droplet method with microbridges (Hampton) placed inside multiwell VDX crystallization plates (Hampton) together with 1 ml reservoir solution and 10 µl top solution. The reservoir solution contained 35% MPD (v/v), 5% PEG 350 monomethylether (MME) (v/v), 25 mM Tris-HCl, pH 7.0, 70 mM SrCl₂ and 250 mM urea. The top solution contained 24.5% MPD (v/v), 3.5% PEG350 (MME) (v/v), 17.5 mM Tris, pH 7.0, 49 mM SrCl₂, 175 mM urea, 3.75 mM Sarkosyl and 10 mg ml⁻¹ protein. Typically, small diamond-shaped crystals appeared overnight and continued to grow to a maximum size of 0.4 mm in the longest dimension within one week. Crystals were cryoprotected by transferring into increasing amounts of MPD, from an initial solution of 55% MPD, 25 mM SrCl₂ and 10 mM Tris, pH 7.0, to a final concentration of 75% MPD. Crystals typically cracked at 65–70% MPD, but reannealed over the course of 30 min. 20 mM DTT was included in the crystallization and cryoprotection solution of the selenomethionine crystals. Cryoprotected crystals were frozen in propane before data collection. Crystals were mosaic (around 1.5°) and showed variability in cell dimensions and diffraction quality (Table 1).

Diffraction data of the native, the selenomethionine-labelled Sn1, and iodine derivatives were collected at beamline 19ID at APS using an in-house CCD detector. The derivative data were processed with HKL2000 (ref. 34) and the native data with d³TREK (Molecular Structure Corp.). MAD data for the Sx, Sn1Sn2, and Sn1Sn2Sx selenomethionine-labelled complexes were collected at beamline 1-5 at SSRL using a Quantum-4 CCD detector (Area Detector Systems) and were processed with DENZO/SCALEPACK³⁴. All diffraction data were collected at 100 K.

Heavy-atom sites. Fifteen of the twenty-one selenium sites of the selenomethionine-labelled-Sn1 data set were found by the Patterson search method implemented in the crystallography and NMR system (CNS)³⁵ using the selenomethionine Sn1 diffraction data at the selenium anomalous peak wavelength. One extra site was found by peak searching using a log-likelihood gradient map³⁶ after MAD phasing. A Fourier difference map showed two iodine substitutions *ortho* to the hydroxyl group in Tyr 88 of Sb. Additional ordered selenomethionine sites from appropriately labelled complexes were obtained by Fourier difference maps using the Sn1Sn2 λ₃ diffraction data. This produced a total of 15 ordered sites out of 18 in Sx, 16 out of 21 in Sn1 and 15 out of 21 in Sn2.

Multi-MAD phasing. Heavy-atom parameter refinement and phasing were carried out individually for the three MAD experiments using CNS (Table 1 in Supplementary Information)³⁵. MAD phasing used the Phillips–Hodgson method³⁷ formulated in a maximum-likelihood framework³⁸. Density modification consisted of solvent flattening³⁹ using an envelope determined by electron-density fluctuations, histogram matching⁴⁰, and phase extension from 3 Å to 2.4 Å resolution using the native diffraction data. Several regions showed only weak or variable electron density, especially at the N-terminal end of the complex, using the individual MAD data sets. Phase combination of the three MAD-phase sets with unit weighting followed by density modification produced a significantly improved electron-density map (Table 1 in Supplementary Information, and Fig. 1). The unweighted phase differences at 50–3.0 Å resolution between the final model and the various phase sets after density modification were: 58.2° for Sx, 48.1° for Sn1Sn2, 51.2° for Sn1SnSx, 43.1° for the combination of Sx and Sn1Sn2, and 42.1° for the combination of all three MAD data sets. The experimental electron-density map showed most of the

model. Most of the remaining parts of the model became traceable in phase-combined electron density maps.

Refinement. The initial model was built using the program O (ref. 41). Refinement was carried out using torsion-angle simulated annealing⁴² and conjugate gradient minimization, interspersed with restrained individual atomic thermal factor refinement⁴³ as implemented in CNS³⁵. The MLHL maximum likelihood target⁴⁴ was used with density-modified phases as prior experimental phase information. Model quality was periodically checked using CNS³⁵. Subsequent model-building stages used phase-combined σ_A -weighted electron density maps⁴⁵ using the combined phases from all three MAD experiments and the present model phases. Refinement included a uniform bulk solvent correction ($B_{\text{sol}} = 38.5 \text{ \AA}^2$; $k_{\text{sol}} = 0.39 \text{ e}^-/\text{\AA}^3$) and overall anisotropic thermal factor ($B_{11} = -1.54 \text{ \AA}^2$; $B_{22} = -8.86 \text{ \AA}^2$; $B_{33} = 10.40 \text{ \AA}^2$) refinement. As the three complexes exhibited significant differences, especially at their N and C termini (Fig. 2b), non-crystallographic symmetry (NCS) restraints were not applied and the complexes were built independently. At this resolution, refinement in the absence of NCS restraints can be justified as high-quality phase information was used in the refinement target function. All diffraction data were used throughout the refinement except that a 10% randomly selected test set for cross-validation σ_A values was used in the maximum likelihood calculations⁴⁷ and calculation of R_{free} (ref. 46). Strontium sites were identified by the average anomalous signal of the ion at the Sx, Sn1Sn2 and Sn1Sn2Sx low-energy-remote (λ_1) wavelengths (Table 1) and by proper divalent-cation-coordinating ligation. Water molecules were placed at sites corresponding to σ_A -weighted-difference Fourier peaks greater than 3σ that exhibited reasonable protein–solvent hydrogen-bonding distances without steric conflict. MPD molecules were also included in the structure on the basis of the overall shape and coordination of the electron density from difference Fourier and σ_A -weighted $2F_o - F_c$ maps. The final model contains residues: Sb 25–93, Sx 188–259, Sn1 7–83 and Sn2 131–204 for complex I; Sb 28–96, Sx 189–261, Sn1 12–83 and Sn2 132–203 for complex II; Sb 26–94, Sx 186–259, Sn1 11–83 and Sn2 132–204 for complex III; 19 water molecules; 13 Sr²⁺ ions; and 7 MPD molecules. The model exhibits excellent stereochemistry and structure quality, with an average bond length and bond-angle deviation of 0.009 Å and 1.34°, respectively, no Ramachandran violations, and a *B*-factor r.m.s.d. for main-chain bonds of 1.38 Å² and for sidechain bonds of 2.4 Å². The overall free *R*-value is 30.3% and the *R* value is 26.5%, using all observed diffraction data between 50 and 2.4 Å resolution. No amplitude-based cutoffs were applied.

Structure analysis. Analysis of α -helical parameters (crossing angle, number of residues per turn, radius of helical bundle and layer analysis) was done for the four-helical-bundle region of the synaptic fusion complex (layers '7' to '8', (Table 2 in Supplementary Information). The crossing angles range between 15° and 25° for synaptobrevin, 5° and 20° for syntaxin, 5° and 20° for Sn1, and 12° and 25° for Sn2. The crossing angles are defined as the dihedral angle between the local helical axes and the bundle axis. The local helical axis is defined as the vector between centroids of the C α , N and C positions of two sets of seven consecutive residues spaced apart by half an α -helical turn. Surface plots and electrostatic analyses were done with GRASP⁴⁸. Figures 1, 2a, 3 and 5 were prepared with gl_render (provided by L. Esser) or Bobscrip⁴⁹, and were rendered using PovRay (<http://www.povray.org/>).

Received 6 August; accepted 21 August 1998.

- Südhof, T. The synaptic vesicle cycle: a cascade of protein–protein interactions. *Nature* **375**, 645–653 (1995).
- Hanson, P. I., Heuser, J. E. & Jahn, R. Neurotransmitter release—four years of SNARE complexes. *Curr. Opin. Neurobiol.* **7**, 310–315 (1997).
- Jahn, R. & Niemann, H. Molecular mechanisms of clostridial neurotoxins. *Ann. NY Acad. Sci.* **733**, 245–255 (1994).
- Söllner, T., Bennett, M. K., Whiteheart, S. W., Scheller, R. H. & Rothman, J. E. A protein assembly-disassembly pathway *in vitro* that may correspond to sequential steps of synaptic vesicle docking, activation, and fusion. *Cell* **75**, 409–418 (1993).
- Hayashi, T., Yamasaki, S., Nauenburg, S., Binz, T. & Niemann, H. Disassembly of the reconstituted synaptic vesicle membrane fusion complex *in vitro*. *EMBO J.* **14**(10), 2317–2325 (1995).
- Fasshauer, D., Otto, H., Eliason, W. K., Jahn, R. & Brunger, A. T. Structural changes are associated with SNARE-complex formation. *J. Biol. Chem.* **272**, 28036–28041 (1997).
- Nichols, B. J., Ungerman, C., Pelham, H. R. B., Wickner, W. T. & Haas, A. Homotypic vacuolar fusion mediated by *t*- and *v*-SNAREs. *Proc. Natl Acad. Sci. USA* **94**, 199–202 (1997).
- Otto, H., Hanson, P. I. & Jahn, R. Assembly and disassembly of a ternary complex of synaptobrevin, syntaxin, and SNAP-25 in the membrane of synaptic vesicles. *Proc. Natl Acad. Sci. USA* **94**, 6197–6201 (1997).
- Block, M. R., Glick, B. S., Wilcox, C. A., Wieland, F. T. & Rothman, J. E. Purification of an N-

- ethylmaleimide-sensitive protein catalyzing vesicular transport. *Proc. Natl Acad. Sci. USA* **85**, 7852–7856 (1988).
10. Hanson, P. I., Otto, H., Barton, N. & Jahn, R. The N-ethylmaleimide-sensitive fusion protein and α -SNAP induce a conformational change in syntaxin. *J. Biol. Chem.* **270**, 16955–16961 (1995).
11. Fasshauer, D., Eliason, W. K., Brunger, A. T. & Jahn, R. Identification of a minimal core of the synaptic SNARE-complex sufficient for reversible assembly and disassembly. *Biochemistry* **37**, 10345–10353 (1998).
12. Hendrickson, W. A. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254**, 51–58 (1991).
13. Hanson, P. I., Roth, R., Morisaki, H., Jahn, R. & Heuser, J. E. Structure and conformational changes in NSF and its membrane receptor complex visualized by quick-freeze/deep-etch electron microscopy. *Cell* **90**, 523–525 (1997).
14. Oyler, G. A. *et al.* The identification of a novel synaptosomal-associated protein, SNAP-25, differentially expressed by neuronal subpopulations. *J. Cell. Biol.* **109**, 3039–3052 (1989).
15. Harbury, P. B., Zhang, T., Kim, P. S. & Alber, T. A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science* **262**, 1401–1407 (1993).
16. Lupas, A., van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252**, 1162–1164 (1991).
17. Crick, F. H. C. The packing of α -helices: simple coiled coils. *Acta Crystallogr.* **6**, 689–697 (1953).
18. Wolf, E., Kim, P. S. & Berger, B. MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci.* **6**, 1179–1189 (1997).
19. Weimbs, T., Mostov, K. E., Low, S. H. & Hofmann, K. A model for structural similarity between different SNARE complexes based on sequence relationships. *Trends Cell Biol.* **8**, 260–262 (1998).
20. Hao, J. C., Salem, N., Peng, X. R., Kelly, R. B. & Bennett, M. K. Effect of mutations in vesicle-associated membrane protein (VAMP) on the assembly of multimeric protein complexes. *J. Neurosci.* **17**, 1596–1603 (1997).
21. Saifee, O., Wei, L. & Nonet, M. L. The *Caenorhabditis elegans* unc-64 locus encodes a syntaxin that interacts genetically with synaptobrevin. *Mol. Biol. Cell* **9**, 1235–1239 (1998).
22. Fasshauer, D., Bruns, D., Shen, B., Jahn, R. & Brunger, A. T. A structural change occurs upon binding of syntaxin to SNAP-25. *J. Biol. Chem.* **272**, 4582–4590 (1997).
23. Kee, Y., Lin, R. C., Hsu, S. C. & Scheller, R. H. Distinct domains of syntaxin are required for synaptic vesicle fusion complex formation and dissociation. *Neuron* **14**, 991–998 (1995).
24. Hayashi, T. *et al.* Synaptic vesicle membrane fusion complex: action of clostridial neurotoxins on assembly. *EMBO J.* **13**, 5051–5061 (1994).
25. Wimley, W. C. & White, S. H. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nature Struct. Biol.* **3**, 842–848 (1996).
26. Chan, D. C., Fass, D., Berger, J. M. & Kim, P. S. Core structure of gp41 from the HIV envelope glycoprotein. *Cell* **89**, 263–273 (1997).
27. Weissenhorn, W., Dessen, A., Harrison, S. C., Skehel, J. J. & Wiley, D. C. Atomic structure of the ectodomain from HIV-1 gp41. *Nature* **387**, 426–430 (1997).
28. Caffrey, M. *et al.* Three-dimensional solution structure of the 44 kDa ectodomain of SIV gp41. *EMBO J.* **17**, 4572–4584 (1998).
29. Wilson, I. A., Skehel, J. J. & Wiley, D. C. Structure of the haemagglutinin membrane glycoprotein influenza virus at 3 Å resolution. *Nature* **289**, 366–373 (1981).
30. Rice, L. M., Brennwald, P. & Brunger, A. T. Formation of a yeast SNARE complex is accompanied by significant structural changes. *FEBS Lett.* **415**, 49–55 (1997).
31. Chan, D. C. & Kim, P. S. HIV entry and its inhibition. *Cell* **93**, 681–684 (1998).
32. Bernard, A. & Payton, M. Fermentation and growth of *Escherichia coli* for optimal protein production. *Curr. Protoc. Protein Sci.* **5.3**, 1–18 (1995).
33. Leahy, D. J., Erickson, H. P., Aukhil, I., Joshi, P. & Hendrickson, W. A. Crystallization of a fragment of human fibronectin: introduction of methionine by site-directed mutagenesis to allow phasing via selenomethionine. *Proteins* **19**, 48–54 (1994).
34. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1998).
35. Brunger, A. T. *et al.* Crystallography & NMR system (CNS): a new software system for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
36. Bricogne, G. Bayesian statistical viewpoint on structure determination: basic concepts and examples. *Methods Enzymol.* **276**, 361–423 (1997).
37. Phillips, J. C. & Hodgson, K. O. The use of anomalous scattering effects to phase diffraction patterns from macromolecules. *Acta Crystallogr. A* **36**, 856–864 (1980).
38. Burling, F. T., Weis, W. I., Flaherty, K. M. & Brunger, A. T. Direct observation of protein solvation and discrete disorder with experimental crystallographic phases. *Science* **271**, 72–77 (1996).
39. Wang, B.-C. Resolution of phase ambiguity in macromolecular crystallography. *Methods Enzymol.* **115**, 90–112 (1985).
40. Zhang, K. Y. J. & Main, P. Histogram matching as a new density modification technique for phase refinement and extension of protein molecules. *Acta Crystallogr. A* **46**, 41–46 (1990).
41. Jones, T. A., Zou, J. Y., Cowan, S. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
42. Rice, L. M. & Brunger, A. T. Torsion angle dynamics: reduced variable conformational sampling enhances crystallographic structure refinement. *Proteins* **19**, 277–290 (1994).
43. Hendrickson, W. A. Stereochemically restrained refinement of macromolecular structures. *Methods Enzymol.* **115**, 252–270 (1985).
44. Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. Incorporation of prior phase information strengthens maximum likelihood structural refinement. *Acta Crystallogr. D* (in the press).
45. Read, R. J. Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr. A* **42**, 140–149 (1986).
46. Brunger, A. T. The free *R* value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–474 (1992).
47. Read, R. J. Model phases: probabilities and bias. *Methods Enzymol.* **277**, 110–128 (1997).
48. Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296 (1991).
49. Esnouf, M. An extensively modified version of MOLSCRIPT that includes greatly enhanced coloring capabilities. *J. Mol. Graph. Model* **15**, 132–134 (1997).
50. Jorgensen, W. L. & Rives, J. T. The OPLS potential functions for protein energy minimizations for crystals of cyclic peptide and crambin. *J. Am. Chem. Soc.* **110**, 1657–1666 (1988).
51. Malashkevich, V. N., Chan, D. C., Chutkowski, C. T. & Kim, P. Crystal structure of the simian immunodeficiency virus (SIV) gp41 core: Conserved helical interactions underlie the broad inhibitory activity of gp41 peptides. *Proc. Natl Acad. Sci. USA* **95**, 9134–9139 (1998).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements. We thank P. D. Adams and C. Ostermeier for advice; D. M. Engelman, K. M. Fiebig, T. Simonson and R. C. Yu for stimulating discussions; H. Bellamy for assistance with data collection at SSRL 1-5 (SSRL is funded by the Department of Energy; the SSRL Biotechnology Program is supported by the NIH; further SSRL support is provided by the Department of Energy); A. Joachimiak and staff at the Structural Biology Center at the Advanced Photon Source for assistance with data collection at 19ID (this national user facility is supported by the Department of Energy); L. Esser for assistance in figure preparation; and J. Pflugrath for help with dTREK. This work was supported by the NIH (A.T.B.).

Correspondence and requests for materials should be addressed to A.T.B. (e-mail: brunger@laplace.csb.yale.edu). The refined coordinates and diffraction data have been deposited with the Brookhaven Protein Databank, accession number 1scf. Coordinates are immediately available from <http://atb.csb.yale.edu>.

Compact protoplanetary disks around the stars of a young binary system

L. F. Rodríguez*, P. D'Alessio*, D. J. Wilner†, P. T. P. Ho‡, J. M. Torrelles§, S. Curiel*, Y. Gómez*, S. Lizano*, A. Pedlar§, J. Cantó* & A. C. Raga*

* Instituto de Astronomía, UNAM Apdo. Postal 70-264, México DF 04510, México

† Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA

‡ Instituto de Astrofísica de Andalucía, CSIC Sancho Panza s/n, Apdo. Correos 3004, E-18080 Granada, Spain, and IMDEA (CSIC-UIB), 07071 Palma, Spain

§ NRAL, University of Manchester, Jodrell Bank, Macclesfield, Cheshire SK11 9DL, UK

Planet formation is believed to occur in the disks of gas and dust that surround young solar-type stars¹. Most stars, however, form in multiple systems^{2–5}, where the presence of a close companion could affect the structure of the disk^{6–8} and perhaps interfere with planet formation. It has been difficult to investigate this because of the resolution needed. Here we report interferometric observations (at a wavelength of 7 mm) of the core of the star-forming region L1551. We have achieved a linear resolution of seven astronomical units (less than the diameter of Jupiter's orbit). The core of L1551 contains two distinct disks, with a separation of 45 AU; these appear to be associated with a binary system. Both disks are spatially resolved, with semi-major axes of about 10 AU, which is about a factor of ten smaller than disks around isolated stars^{9–12}. The disk masses are of order 0.05 solar masses, which could be enough to form planetary systems like our own.

L1551 is a molecular cloud in Taurus that is known to be undergoing intensive star formation, although restricted to low-mass stars (that is, with masses comparable to or smaller than that of the Sun). The embedded infrared source L1551 IRS5 (ref. 13) is believed to be associated with a very young stellar source. Its bolometric luminosity is ~ 30 solar luminosities¹⁴ and it is embedded in a dense envelope of molecular gas and dust that extends over $\sim 10^3$ – 10^4 AU (refs 15, 16). The continuum millimetre emission, when observed with angular resolutions of ~ 1 arcsec, has been interpreted as originating from heated dust in a protoplanetary disk with dimensions of about 100 AU (refs 17, 18).

In the centimetre wavelengths, there is compact radio continuum emission with elongated morphology^{19,20} that aligns with the large-scale bipolar molecular outflow²¹. This centimetre radiation is known to be free-free emission that originates in the ionized outflowing gas²². However, when observed with subarc second angular resolution, the core of the centimetre emission from L1551 IRS5 breaks into two compact components separated by 0.3 arcsec, which have been interpreted as being either a protobinary system¹⁹ or the inner ionized edges of a gas and dust toroid around a single star²⁰. As the millimetre continuum emission, tracing the dust, has been interpreted as coming from a single disk, the single-star interpretation has been favoured recently. However, there are several observations that suggest that two independent outflow systems emanate from L1551 IRS5 (refs 23–25), favouring a binary nature for the source. Furthermore, observations at 2.7-mm (ref. 26) with angular resolution of about 0.5 arcsec have been used to argue for the presence of two unresolved disks in L1551 IRS5.

Our observations were made with the Very Large Array of the National Radio Astronomy Observatory in its highest angular resolution A configuration. The 7-mm observations were made in 1997 January 10, and the 3.6-cm observations were made in 1996 December 10. As the observations are separated by only one month,

we can consider them to be practically simultaneous. The data were edited, calibrated and imaged using the program AIPS of NRAO.

The 7-mm map is shown in Fig. 1: two compact sources are clearly detected and their parameters are given in the legend to Fig. 1. Our angular resolution is an order of magnitude better than previous millimetre observations²⁶, and clearly resolves the two components. We believe that the 7-mm emission from L1551 IRS5 traces dust in protoplanetary disks based on two main arguments. First, these millimetre sources are elongated approximately in the north–south direction, whereas the centimetre emission, which is believed to trace the ionized jets and is shown in Fig. 2, is elongated approximately in the east–west direction. Disk and jet are expected to show this relative perpendicular orientation. Second, the 7-mm flux densities of these two compact components are too large to be accounted for as being due to free-free radiation, and are most probably produced by heated dust in protoplanetary disks. Indeed, the spectra of these two components (see Table 1 and Fig. 3) shows that the 7-mm data points fall significantly above the extrapolation of the centimetre flux densities and that the emission originates mostly from heated dust. To our knowledge, this is the first time that protoplanetary disks have been imaged with a resolution below 10 AU.

Our data then indicate that L1551 IRS5 is a binary system with 45 AU separation, as first proposed more than a decade ago¹⁹ and

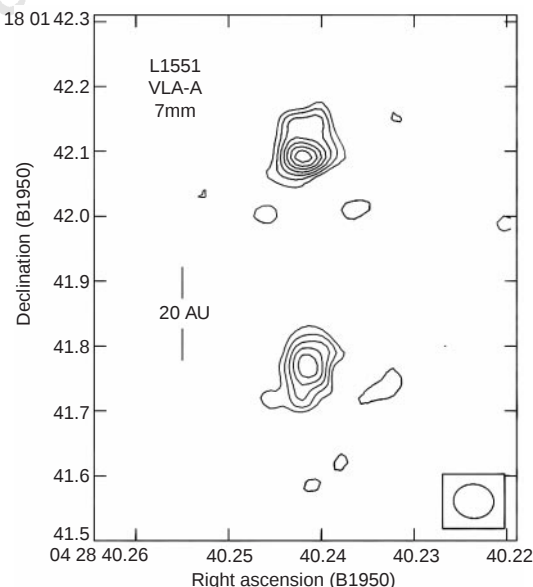


Figure 1 Cleaned, natural-weight Very Large Array (VLA) map of the L1551 IRS5 region at 7 mm. Two compact sources are evident in the map. Contours are $-4, -3, 3, 4, 5, 6, 7, 8$ and 9 times $0.24 \text{ mJy beam}^{-1}$, the r.m.s. noise of the map. The half-power contour of the beam (with dimensions of $0.062'' \times 0.052''$ and position angle of 85°) is shown in the bottom right corner. A scale of 20 AU is also shown. The emission shown in this map is interpreted to arise from two compact protoplanetary structures forming a gravitationally bound binary system. The peak positions, total flux densities, and deconvolved angular dimensions of the sources are $\alpha(1950) = 04 \text{ h } 28 \text{ m } 40.242 \text{ s}$, $\delta(1950) = +18^\circ 01' 42.09''$, $7.4 \pm 0.5 \text{ mJy}$, and $0.12'' \times 0.08''$; 159° for source A (northern), and $\alpha(1950) = 04 \text{ h } 28 \text{ m } 40.242 \text{ s}$, $\delta(1950) = +18^\circ 01' 41.77''$, $4.8 \pm 0.5 \text{ mJy}$, and $0.10'' \times 0.05''$; 171° for source B (southern). The positional error is estimated to be $\sim 0.03''$; the errors in deconvolved angular size and position angle are $\sim 0.01''$ and $\sim 10^\circ$, respectively. These 7-mm observations were made using the fast-switching technique between source and phase calibrator with 2-min cycles. Although very demanding in time overhead, this technique is known to correct for the phase fluctuations present in interferometer millimetre observations. Our results indicate that, under good weather conditions, the correction is effective even at the 30-km baselines of these observations. As primary amplitude calibrator we used 1328 + 307, with adopted flux density of 1.45 Jy , and as phase calibrator we used 0428 + 205, with bootstrapped flux density of $0.41 \pm 0.01 \text{ Jy}$.

further implied by recent millimetre studies²⁶. Furthermore, we find that each of the components of the binary system is surrounded by a protoplanetary disk with a semi-major axis of about 10 AU. Analysis of the available millimetre data²⁶ suggests that, in addition to the compact disks, there is a circumbinary structure and a large-scale envelope around L1551 IRS5. However, our high-angular-resolution observations are not sensitive to these extended components, although the 7-mm map shows some evidence of faint, extended emission around the compact sources.

The average brightness temperature of the 7-mm sources is quite high, about 400 K. This suggests that these small disks are relatively hot and optically thick, even at 7 mm. The observed parameters of the 7-mm sources, plus the constraints of the bolometric luminosity and the radio continuum flux densities, can be modelled as viscous disks^{27,28}. We have calculated time-independent, geometrically thin, viscous accretion disk models that are heated by the viscous dissipation in the disk itself, neglecting the irradiation of the central star which is much less important in the energy budget. The disk is assumed to be in vertical hydrostatic equilibrium and the energy is transported by radiative, convective and turbulent fluxes. The surface brightness distribution of the disk is calculated by solving the transfer equation along rays towards the observer. These disks are hotter in their interior than in their surfaces. Then, millimetre observations that penetrate deeper into the disk (because of the lower opacity of dust at these wavelengths) 'see' hotter gas than infrared observations that do not penetrate as deeply. The best fits to the data are obtained for disks with radii of ~ 10 AU, total masses of $\sim 0.06 M_{\odot}$ (north component) and $\sim 0.03 M_{\odot}$ (south component), and mass accretion rates of $\sim 5 \times 10^{-6} M_{\odot} \text{ yr}^{-1}$ (Fig. 3).

There is the possibility that the 7-mm flux density is contaminated by free-free emission from ionized outflowing gas. As both the derived mass and accretion rate for the disk depend to a first approximation linearly on the 7-mm flux density, free-free contamination of 25, 50 and 75% would result in decreases of the order of 1.3, 2 and 4, respectively, in the derived mass and accretion rate. Even in the case of the largest free-free contamination considered, the derived masses for the disks remain at values comparable with the minimum mass of $\sim 0.01 M_{\odot}$ that is required to form a planetary system similar to our own.

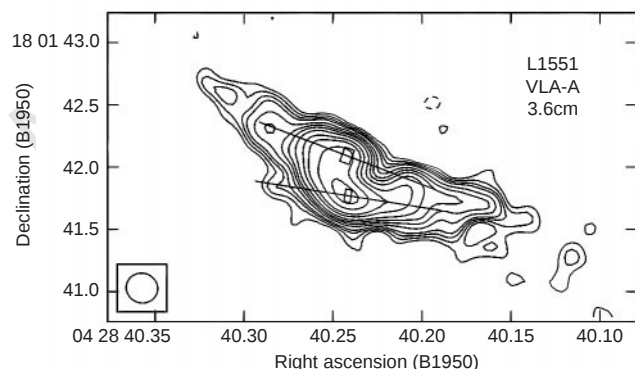


Figure 2 Cleaned VLA map of the L1551 IRS5 region at 3.6 cm made with the task IMAGR of AIPS and Briggs' ROBUST parameter set equal to -0.5 . Contours are $-4, -3, 3, 4, 5, 6, 8, 10, 12, 15, 20, 25, 35$ and 45 times $11 \mu\text{Jy beam}^{-1}$, the r.m.s. noise of the map. The half-power contour of the beam (with dimensions of $0.25'' \times 0.24''$ and position angle of 73°) is shown in the bottom left corner. The emission shown in this map originates from ionized, outflowing gas. The two rectangles shown at the centre of the map mark the positions, deconvolved dimensions, and orientations of the 7-mm sources shown in Fig. 1. The solid lines are drawn perpendicular to the major axes of the 7-mm sources, and may delineate the axes of two independent jets centred on the two 7-mm sources. This appears to be consistent with the asymmetric morphology of the 3.6-cm emission. As amplitude calibrator we used 1328 + 307, with adopted flux density of 3.50 Jy ; as phase calibrator we used 0507 + 179, with bootstrapped flux density of 0.916 ± 0.001 .

We conclude that these disks have sufficient mass to produce planets. However, their relatively small size could inhibit the formation of outer, icy planets similar to Uranus and Neptune. Furthermore, the truncation observed in these disks suggests that in very close binaries (that is, with separation of a few astronomical units), planet formation could be inhibited. Our results also indicate that the collimation of the powerful outflows characteristically observed in young stars can be achieved in a disk of compact dimensions.

The two 7-mm sources are separated in the sky by $0.32'' \pm 0.02''$ ($45 \pm 3 \text{ AU}$), with a position angle of $180 \pm 4^\circ$. This separation is close to the average major axis observed in pre-main²⁹ and main sequence³⁰ binaries. If these sources are indeed tracing the components of a binary system with solar-mass stars, we expect its period to be of the order of 100 years and that significant proper motions should be detected in a decade. It is also interesting that both disks have similar dimensions and masses to within a factor of two. It has been suggested³¹ that in a binary system the individual disks could

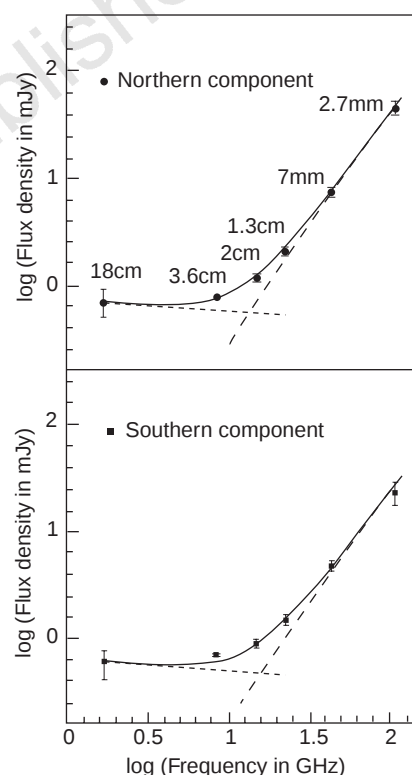


Figure 3 Centimetre and millimetre spectrum of the northern (top) and southern (bottom) components in L1551 IRS5 (data from Table 1). The low-frequency fluxes are believed to be dominated by free-free emission from ionized outflows (like those shown in Fig. 2) and have been modelled (short-dashed line) as being due to optically thin free-free emission, whereas the high-frequency fluxes most probably originate from heated dust in a viscous protoplanetary disk and have been modelled (long-dashed line) as described in the text. These models use the standard α -viscosity prescription that provides a value for the viscosity parameter in terms of the local sound speed and pressure scale height of the gas. In our models, we adopt $\alpha \approx 0.02$. A mass of 0.5 solar masses is adopted for the stars in the binary system. The opacity for the disk models includes contributions from dust and gas²⁸. The solid line is the addition of the ionized gas and dust contributions. The 7-mm emission appears to be dominated by the dust component. The compact disk models (P. D'Alessio *et al.*, manuscript in preparation) have been projected in the sky, convolved with our observational beam and finally fitted with gaussian ellipsoids until good agreement was obtained with the observations. The best fits for the inclination angle (the angle between the line of sight and the symmetry axis of the disk) are 50° and 60° for the northern and southern components, respectively.

Table 1 Integrated flux densities for the two components of L1551 IRS5

Wavelength	A (North)	B (South)	Interferometer	Reference
18 cm	0.70 ± 0.20	0.60 ± 0.20	MERLIN	*
3.6 cm	0.78 ± 0.02	0.70 ± 0.02	VLA	This work
2.0 cm	1.2 ± 0.1	0.9 ± 0.1	VLA	20
1.3 cm	2.0 ± 0.2	1.5 ± 0.2	VLA	†
7.0 mm	7.4 ± 0.5	4.8 ± 0.5	VLA	This work
2.7 mm	45 ± 6	23 ± 6	BIMA	26

Flux densities are in mJy.

* S. Curiel *et al.*, manuscript in preparation.

† D. W. Koerner and A. I. Sargent, manuscript in preparation.

still accrete from the circumbinary disk, via gas streams that move across the tidal gap. Under these circumstances, accretion will occur preferentially into the lower mass component of the system, driving the mass ratio towards unity, as observed in L1551 IRS5.

In the case of the visible T Tauri stars, which are sources more evolved than L1551 IRS5, there is evidence^{7,8} that the stars constituting a close binary (that is, a pair with separation of less than 100 AU) have less millimetre emission than single stars or wide binaries. This result has been taken to indicate that the gravitational interactions between the members of the binary result in less massive disks. In the case of extremely young stars like L1551 IRS5, there are no data available yet to search for similar tendencies. □

Received 19 March; accepted 26 June 1998.

- Beckwith, S. V. W. & Sargent, A. I. Circumstellar disks and the search for neighbouring planetary systems. *Nature* **383**, 139–144 (1996).
- Reipurth, B. & Zinnecker, H. Visual binaries among pre-main sequence stars. *Astron. Astrophys.* **278**, 81–108 (1993).
- Leinert, C. *et al.* A systematic approach for young binaries in Taurus. *Astron. Astrophys.* **278**, 129–149 (1993).
- Simon, M. *et al.* A lunar occultation and direct imaging survey of multiplicity in the Ophiuchus and Taurus star-forming regions. *Astrophys. J.* **443**, 625–637 (1995).
- Ghez, A. M., McCarthy, D. W., Patience, J. L. & Beck, T. L. The multiplicity of pre-main-sequence stars in southern star-forming regions. *Astrophys. J.* **481**, 378–385 (1997).
- Artymowicz, P. & Lubow, S. H. Dynamics of binary-disk interaction. I: Resonances and disk gap sizes. *Astrophys. J.* **421**, 651–667 (1994).
- Jensen, E. L. N., Mathieu, R. D. & Fuller, G. A. The connection between submillimetre continuum flux and binary separation in young binaries: Evidence of interaction between stars and disks. *Astrophys. J.* **458**, 312–326 (1996).
- Osterloh, M. & Beckwith, S. V. W. Millimeter-wave continuum measurements of young stars. *Astrophys. J.* **439**, 288–302 (1995).
- Lay, O. P., Carlstrom, J. E., Hills, R. E. & Phillips, T. G. Protostellar accretion disks resolved with the JCMT-CSO interferometer. *Astrophys. J.* **434**, L75–L78 (1994).
- Mundy, L. G. *et al.* Imaging the HL Tauri disk at $\lambda = 2.7$ millimeters with the BIMA array. *Astrophys. J.* **464**, L169–L173 (1996).
- Wilner, D. J., Ho, P. T. P. & Rodriguez, L. F. Subarcsecond VLA observations of HL Tauri: Imaging the circumstellar disk. *Astrophys. J.* **470**, L117–L121 (1996).
- Mannings, V., Koerner, D. W. & Sargent, A. I. A rotating disk of gas and dust around a young counterpart to β Pictoris. *Nature* **388**, 555–557 (1997).
- Strom, K. M., Strom, S. E. & Vrba, F. J. Infrared surveys of dark-cloud complexes. IV. The Lynds 1517 and Lynds 1551 clouds. *Astron. J.* **81**, 320–322 (1976).
- Butner, H. M., Evans, N. J., Lester, D. F., Levreault, R. M. & Strom, S. E. Testing models of low-mass star formation—High-resolution far-infrared observations of L1551 IRS 5. *Astrophys. J.* **376**, 635–653 (1991).
- Torrelles, J. M. *et al.* Are interstellar toroids the focusing agent of the bipolar molecular outflows? *Astrophys. J.* **274**, 214–230 (1983).
- Saito, M., Kawabe, R., Kitamura, Y. & Sunada, K. Imaging of an infalling disklike envelope around L1551 IRS 5. *Astrophys. J.* **473**, 464–469 (1996).
- Keene, J. & Masson, C. R. Detection of A 45 AU radius source around L1551-IRS5—A possible accretion disk. *Astrophys. J.* **355**, 635–644 (1990).
- Lay, O. P., Carlstrom, J. E., Hills, R. E. & Phillips, T. G. Protostellar accretion disks resolved with the JCMT-CSO interferometer. *Astrophys. J.* **434**, L75–L78 (1994).
- Biegging, J. H. & Cohen, M. Multifrequency radio images of L1551 IRS 5. *Astrophys. J.* **289**, L5–L8 (1985).
- Rodriguez, L. F., Cantó, J., Torrelles, J. M. & Ho, P. T. P. The double radio source associated with L1551 IRS 5—Binary system or ionized circumstellar torus? *Astrophys. J.* **301**, L25–L28 (1986).
- Snell, R. L., Loren, R. B. & Plambeck, R. L. Observations of CO in L1551—Evidence for stellar wind driven shocks. *Astrophys. J.* **239**, L17–L22 (1980).
- Rodriguez, L. F. in *Herbig-Haro Flows and the Birth of Low Mass Stars* (eds Reipurth, B. & Bertout, C.) 83–92 (Proc. IAU Symp. 182, Kluwer, Dordrecht, 1997).
- Moriarty-Schieven, G. H. & Wannier, P. G. A second outflow from L1551/IRS 5? *Astrophys. J.* **373**, L23–L26 (1991).
- Pound, M. W. & Bally, J. Two new molecular outflows in L1551? *Astrophys. J.* **383**, 705–713 (1991).
- Fridlund, C. V. M., Hultgren, M. & Liseau, R. in *Herbig-Haro Flows and the Birth of Low Mass Stars* (eds Reipurth, B. & Bertout, C.) 19–28 (Proc. IAU Symp. 182, Kluwer, Dordrecht, 1997).
- Looney, L. W., Mundy, L. G. & Welch, W. J. High-resolution $\lambda = 2.7$ millimeter observations of L1551 IRS 5. *Astrophys. J.* **484**, L157–L160 (1997).
- Boss, A. P. Evolution of the Solar Nebula. III. Protoplanetary disks undergoing mass accretion. *Astrophys. J.* **469**, 906–920 (1996).
- D'Alessio, P., Calvet, N. & Hartmann, L. The structure and emission of accretion disks irradiated by infalling envelopes. *Astrophys. J.* **474**, 397–406 (1997).
- Mathieu, R. D. Pre-main-sequence binary stars. *Annu. Rev. Astron. Astrophys.* **32**, 465–530 (1994).

- Duquennoy, A. & Mayor, M. Multiplicity among solar-type stars in the solar neighbourhood. II—Distribution of the orbital elements in an unbiased sample. *Astron. Astrophys.* **248**, 485–524 (1991).
- Artymowicz, P. & Lubow, S. H. Mass flow through gaps in circumbinary disks. *Astrophys. J.* **467**, L77–L80 (1996).

Acknowledgements. We appreciate the open policy of the National Radio Astronomy Observatory, which is a facility of the US National Science Foundation operated under cooperative agreement by Associated Universities, Inc. In particular, we thank the NRAO personnel involved in the 7-mm upgrade. The 7-mm receivers used for these observations were built with the support of CONACyT, México. We thank A. Poveda for his comments on binary stars.

Correspondence should be addressed to L.F.R. (e-mail: luisfr@astro.unam.mx).

Continuous excitation of planetary free oscillations by atmospheric disturbances

Naoki Kobayashi* & Kiwamu Nishida*†

* Earth and Planetary Sciences, Tokyo Institute of Technology, Meguro, Tokyo 152, Japan

Seismology provides a powerful tool for probing planetary interiors^{1,2}, but it has been considered inapplicable to tectonically inactive planets where earthquakes are absent. Here, however, we show that the atmospheres of solid planets are capable of exerting dynamic pressure on their surfaces, thereby exciting free oscillations with amplitudes large enough to be detected by modern broad-band seismographs. Order-of-magnitude estimates of these forces give similar amplitudes of a few nanogals for the Earth, Venus and Mars despite widely varying atmospheric and ambient conditions. The amplitudes are also predicted to have a weak frequency dependence. Our analysis of seismograms, recorded continuously from 1992 to 1993 at 13 globally distributed stations, shows strong evidence for continuously excited fundamental-mode free oscillations on the Earth. This result, together with other recent studies^{3–5}, is consistent with our estimate of atmospheric forcing and we therefore propose that it may be possible to detect atmospheric excitation of free oscillations on Venus and Mars as well.

The Sun pulsates owing to acoustic waves generated by intensive turbulent gaseous motion on the top of the convective layer⁶. Similar processes should occur on solid planets having atmospheres, and should excite seismic waves that propagate throughout the solid interior, thereby inducing free oscillations. Such planetary atmospheric disturbances should occur even on tectonically inactive solid planets such as Mars and Venus.

We now estimate the order of magnitude of the seismic free oscillations excited by atmospheric disturbances. Some atmospheric gases such as water vapour and carbon dioxide cause a so-called greenhouse effect⁷. Atmospheres are relatively transparent to visible light but have strong absorption at infrared wavelengths⁸. Visible light absorbed at and near the surface is converted and re-radiated at infrared wavelengths. The infrared rays are absorbed into the atmosphere and warm the air near the surface. The resultant buoyant force induces atmospheric convective motions. In a steady state, the heat flux transported by the atmosphere must be balanced by the net solar influx:

$$\frac{v^2}{H} = g\alpha\Delta T \quad (1)$$

$$S(1 - A)/4 = \rho_{\text{at}} C_p \Delta T v \quad (2)$$

Symbols used are defined in Table 1. From the above equations, the amplitude of the velocity fluctuation is:

$$v = (g\alpha S(1 - A)/4\rho_{\text{at}} C_p)^{1/3} \quad (3)$$

† Present address: Earthquake Research Institute, University of Tokyo, Tokyo, Japan.

The associated dynamic pressure fluctuation is scaled by $p_0 = \rho_{\text{at}} v^2$, and pressure disturbances with timescale $1/f$ (where f is frequency) are:

$$\delta p = p_0 \frac{f_0}{f} \quad (4)$$

where $1/f_0 = H/v$ is the timescale of convective motion. This frequency dependence is obtained from dimensional considerations. To construct the dimension of pressure in terms of S , H and time $1/f$, we require S/Hf , to which the right-hand side of equation (4) is proportional. Although equation (4) was constructed by dimensional analysis, it is consistent with observations of atmospheric pressure. Here we consider only temporal variations of convective cellular motions whose scale is about H . Smaller-scale turbulence may be important for generation of acoustic waves in the atmosphere, but is less effective in generating elastic waves in the solid portion of the planet.

Fluctuations of pressure loading on the surface of the solid portion will preferentially excite fundamental spheroidal modes, because their elastic energy is confined to shallower depths than overtone modes. Here we use a simplified form of the comprehensive theory of excitation of free oscillations⁹ to assess the excited amplitudes of fundamental spheroidal modes. We consider a steady state in which a balance is achieved between energy input from atmosphere to a mode (energy input rate G) and dissipation of its elastic energy E (the dissipation rate ΓE) in the same manner as is done in helioseismology¹⁰:

$$G = \Gamma E = \frac{\omega}{Q} E \quad (5)$$

Here $E \approx \omega^2 M_e d^2/2$, where d is displacement, ω is angular frequency and M_e is effective elastic mass (solid density $\times$ surface area $\times$ penetration depth of the elastic energy). The input power is $G = \omega d F$, where the force exerted on a mode is $F = 4\pi R^2 \delta p/L$. The cut-off angular order is $L = 2\pi R/H$. We assume that the horizontal correlation length of the random atmospheric disturbance is approximately equal to H , and the spatial power of the disturbance is equipartitioned into L^2 modes. From the above equations, we obtain the amplitude of the excited acceleration:

$$a = \omega^2 d = \frac{QH\rho_0 f_0}{\pi R \rho_{\text{sol}} \lambda f} = \frac{QH\rho_0 f_0}{\pi R \rho_{\text{sol}} c} \quad (6)$$

where we use the horizontal wavelength $\lambda = cf$ as the penetration depth. The frequency dependence of this amplitude is weak, as the quality factor Q and the surface-wave phase velocity c are not strongly dependent on frequency in the millihertz band. Using typical values for various parameters, we obtain amplitudes of the order of nanogals ($10^{-11} \text{ m s}^{-2}$) for the Earth, Venus and Mars (Table 1) in spite of greatly differing solar energy input rates, planetary masses and atmospheric densities. Such amplitudes are detectable by standard broad-band seismographs^{11,12}.

We have analysed continuous records from 13 stations of the Incorporated Research Institutions for Seismology (IRIS) network¹³

for 1992–93 to search for continuously excited free oscillations of the Earth. The instruments are STS-1¹¹ very broad-band seismometers. We selected the 13 stations which have the lowest ground noise level (slightly less than $10^{-18} \text{ m}^2 \text{ s}^{-4} \text{ Hz}^{-1}$ in the millihertz band) in the IRIS station catalogue. For each station, we produced a spectrogram (a contour plot of power spectral density in the frequency–time plane) after removing glitches, nonlinear responses due to large earthquakes, and tidal components: we divided the data into segments of about a half day long with overlaps of about a quarter of a day. We then calculated power spectral densities for each segment using Welch tapering, and lined up the spectrograms with respect to time. These spectrograms show intervals with large oscillations excited by earthquakes of various magnitudes.

To exclude signals due to earthquakes with seismic moment $M_0 > 10^{17} \text{ N m}$, we calculated the power $A_5(t_0)$ of segments in the band from 2.5 to 7.5 mHz (based on the Harvard Centroid-Moment Tensor (CMT) catalogue¹⁴) and discarded segments from the origin time until the time when $A_5(t_0) \exp[-f(t - t_0)/Q]$ ($f = 5 \text{ mHz}$ and $Q = 200$) becomes less than $5.0 \times 10^{-22} \text{ m}^2 \text{ s}^{-4}$ (about a tenth of the noise level). We also discarded segments in which the power $A_{10}(t)$ in the band from 7.5 to 12.5 mHz is larger than $5.0 \times 10^{-21} \text{ m}^2 \text{ s}^{-4}$ to exclude the effects of local smaller earthquakes and other transient phenomena. Further details will be presented elsewhere (K.N. and N.K., manuscript in preparation). The criteria for identifying continuous signals that we interpret as atmospherically excited free oscillations are as follows: (1) fundamental spheroidal modes are selectively excited with nanogal acceleration in a broad frequency range, (2) weak frequency dependence of the amplitudes.

Figure 1a shows selected data at SUR, which is one of the quietest stations. Weak but definite vertical stripes at the frequencies of fundamental spheroidal modes are identified. An ensemble average of the power spectral density is shown in Fig. 1b. From 3 to 7 mHz, distinct peaks are observed at the frequencies of the fundamental spheroidal modes. The power spectral densities are about $1 \times 10^{-18} \text{ m}^2 \text{ s}^{-4} \text{ Hz}^{-1}$ and do not show marked dependence on the frequency. To estimate the acceleration, we multiplied these peak values by the full width at half the peak value and took the square root. The obtained values are slightly less than $1 \times 10^{-11} \text{ m s}^{-2}$, that is, 1 nGal. This is consistent with our order-of-magnitude estimate (above) for the amplitude of free oscillations due to continuous atmospheric excitation. We obtained similar results from other stations (Fig. 2).

Continuously excited signals with frequency-independent amplitudes at the nanogal level have also been observed in other studies^{3–5}. These results, taken together, suggest the existence of continuously excited free oscillations in a broad frequency range from 0.3 to 7 mHz. Nawa *et al.*³ reported seasonal variations in amplitude (higher amplitudes in the period from May to September), but we have not yet confirmed seasonal variations in our data. It is necessary to use a longer time series to clearly verify the existence of seasonal variations, because variations of only $\sim 10\%$ are expected.

Table 1 Physical parameters and excited amplitudes

	S (W m^{-2})	A	ρ_{at}^* (kg m^{-3})	T (K)	R (10^6 m)	g (m s^{-2})	H (10^4 m)	C_p ($\text{J K}^{-1} \text{ kg}^{-1}$)	v (m s^{-1})	ρ_0 (Pa)	τ_0 (10^3 s)	$a^\dagger$ (nGal)
Venus	2,620	0.78	65.3	750	6.03	8.9	1.58	657	0.9	48	18	2.2
Earth	1,370	0.30	1.16	290	6.38	9.8	0.87	1,030	3.8	17	2.2	3.4‡
Mars	590	0.16	1.33×10^{-2}	240	3.40	3.7	1.21	657	13	2.6	0.88	3.3

Here S is solar energy flux, A albedo, ρ_{at} density of atmosphere at surface, T surface temperature, R radius of planet, g surface gravity, H pressure scale height, C_p specific heat at constant pressure, v and ρ_0 are scales of velocity and pressure fluctuations, respectively, and $\tau_0 = 1/f_0$ is timescale of convective motions in the atmosphere.

* We treat atmospheres as diatomic ideal gases. Molecular weights are 44, 28 and 44 for Venus, Earth and Mars, respectively.

† We used solid density $\rho_{\text{sol}} = 4 \times 10^3 \text{ (kg m}^{-3}\text{)}$, surface wave velocity $c = 5 \times 10^3 \text{ (m s}^{-1}\text{)}$ and quality factor $Q = 200$ to evaluate the resultant amplitudes. These values are typical for fundamental spheroidal modes in the millihertz band. We also used these values for Venus and Mars.

‡ On the Earth, energy is carried by not only sensible heat but also by latent heat of water vapour. If we take the latent heat in our estimate, the excitation amplitude has the somewhat smaller value of $\sim 0.8 \text{ nGal}$.

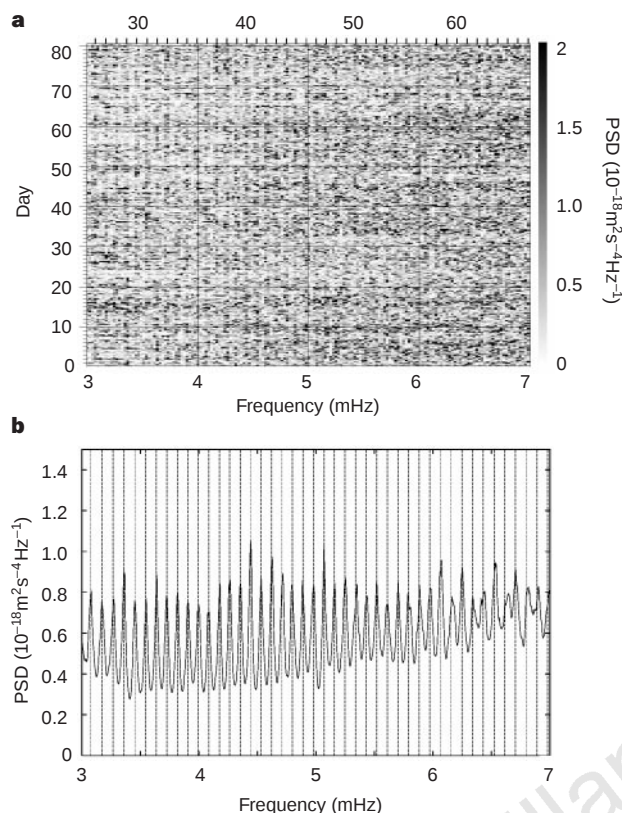


Figure 1 Earth's free oscillations on the quietest days. **a**, A selected spectrogram obtained from the IRIS data at SUR (Southernland, Republic of South Africa) in 1992–93. Grey-scale at right shows spectral power density. Dark vertical stripes can be seen at the frequencies of fundamental spheroidal modes of the spherically symmetric Earth model (PREM)². The level of the stripes is $\sim 1 \times 10^{-18} \text{ m}^2 \text{ s}^{-4} \text{ Hz}^{-1}$. Tick marks on the upper x-axis show the frequencies of spheroidal fundamental modes (from ${}_0S_{22}$ to ${}_0S_{65}$). We cannot see any free oscillation signals outside this frequency range. **b**, Ensemble average of power spectral densities (PSD) of **a**. The distinct peaks correspond to the eigenfrequencies of fundamental spheroidal modes. The vertical lines show locations of the eigenfrequencies of PREM² from ${}_0S_{22}$ to ${}_0S_{65}$. Each peak just overlaps the vertical line. We can roughly estimate the acceleration amplitudes by multiplying peak values by full width at half the peak value and taking the square root. The amplitudes of slightly less than 1 nGal and the weak frequency-dependence are consistent with our estimates.

Although we attempted to exclude the influence of earthquakes in our analysis, we consider further whether the observed continuously excited oscillations might have been driven by earthquakes. Small earthquakes (with seismic moment $M_0 < 10^{17} \text{ N m}$ and not listed in the CMT catalogue) might induce incessant low-amplitude free oscillations. To evaluate the amplitudes of steady-state free oscillations excited by small earthquakes, we calculated the mean seismic moment release rate $\langle \dot{M}_0 \rangle$ using the well known (Gutenberg–Richter) magnitude–frequency relation¹⁵ in the seismic moment range from 0 to 10^{17} N m , and obtained $\langle \dot{M}_0 \rangle = 2 \times 10^{12} \text{ N m s}^{-1}$. Then the amplitudes are written as;

$$a = \frac{2\langle \dot{M}_0 \rangle}{\Gamma \lambda M_e} = \frac{\langle \dot{M}_0 \rangle Q f}{(2\pi c R)^2 \rho_{\text{sol}}}, \quad (7)$$

where a is roughly proportional to f and $a \approx 1 \times 10^{-3} \text{ nGal}$ at 3 mHz. Thus we conclude that the observed continuously excited free oscillations were not excited by earthquakes. Another candidate for the origin of the observed signals is slow earthquakes, which could selectively excite low-frequency modes¹⁶. To examine this

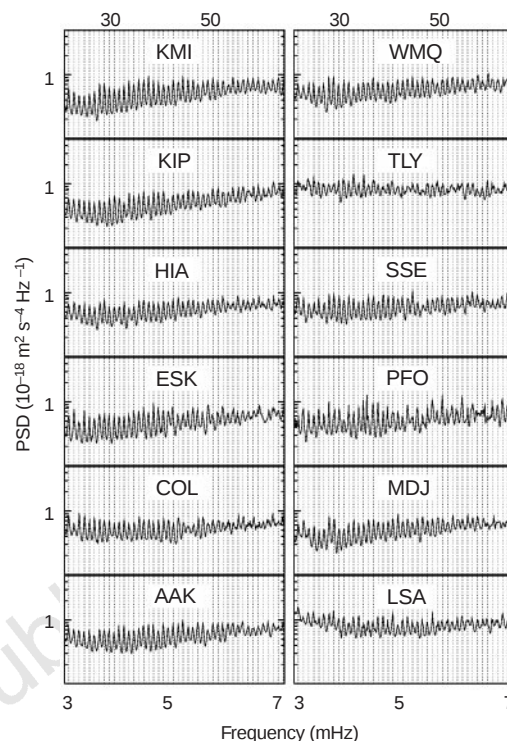


Figure 2 Continuously excited free oscillations were observed globally. Diagrams were produced in the same way as Fig. 1b for 12 different stations: KMI (Kunming, China), KIP (Kipapa, Hawaii, USA), HIA (Hailar, China), EK (Eskdalemuir, UK), COL (College, Alaska, USA), AAK (Ala Archa, Kyrgyzstan), WMQ (Urumqi, China), LY (Talaya, Russia), E (Sheshan, China), PFO (Pinyon Flat, California, USA), MDJ (Mutanjiang, China), LSA (Lhasa, China). The results are similar to those for SUR. The frequencies of the peaks are essentially equal in all of these figures, and they agree with the eigenfrequencies of fundamental spheroidal modes of PREM². This is strong evidence that the observed peaks are truly the Earth's free oscillations because they are observed globally at the same level.

possibility, we carefully searched for abrupt and coherent excitation of low-frequency modes in our spectrograms, but no instance of such signals was found. We also considered the possibility of oceanic excitation but, for the following reasons, this mechanism is less effective than atmospheric excitation. First, convective motions in the ocean are weak mainly because the input energy (mean heat flow from the ocean bottom is $\sim 7.8 \times 10^{-2} \text{ W m}^{-2}$; ref. 17) is lower than the input solar radiation to the atmosphere. Second, microseisms¹⁵ (seismic waves excited by ocean waves induced by winds) have their maximum power near 100 mHz, far from the free-oscillation band.

We conclude that the Earth's atmosphere can continuously excite detectable free oscillations. The amplitudes and frequency dependence of the observed signals are consistent with our theory. Mars and Venus should also have observable continuously excited fundamental spheroidal mode free oscillations excited by their atmospheres; it should be possible to use such signals to probe the internal structure of these planets, at least the upper-mantle structure. We consider that seismological observations by broad-band seismometers on Mars and Venus are likely to be a promising method of probing their interiors. □

Received 14 November 1997; accepted 22 June 1998.

1. Gilbert, F. & Dziewonski, A. M. An application of normal mode theory to the retrieval of structural parameters and source mechanisms from seismic spectra. *Phil. Trans. R. Soc. Lond. A* **278**, 187–269 (1975).
2. Dziewonski, A. M. & Anderson, D. L. Preliminary reference Earth model. *Phys. Earth Planet. Inter.* **25**, 297–356 (1981).
3. Nawa, K. *et al.* Incessant excitation of the Earth's free oscillations. *Earth Planets Space* **50**, 3–8 (1998).
4. Suda, N., Nawa, K. & Fukao, Y. Earth's background free oscillations. *Science* **279**, 2089–2091 (1998).

5. Tanimoto, T., Um, J., Nishida, K. & Kobayashi, N. Earth's continuous oscillations observed on seismically quiet days. *Geophys. Res. Lett.* **25**, 1553–1556 (1998).
6. Goldreich, P. & Kumar, P. The interaction of acoustic radiation with turbulence. *Astrophys. J.* **326**, 462–478 (1988).
7. Sagan, C. Structure of the lower atmosphere of Venus. *Icarus* **1**, 151–169 (1962).
8. Howard, J. N., Burch, D. E. & Williams, D. Infrared transmission of synthetic atmospheres. Parts I–V. *J. Opt. Soc. Am.* **46**, 186–190, 237–241, 242–245, 334–338, 452–455 (1956).
9. Gilbert, F. Excitation of the normal modes of the Earth by earthquake sources. *Geophys. J. R. Astron. Soc.* **22**, 223–226 (1970).
10. Osaki, Y. in *Progress of Seismology of the Sun and Stars* (eds Osaki, Y. & Shibahashi, H.) 75–86 (Lecture Notes in Physics Vol. 367, Springer, Berlin, 1989).
11. Wielandt, E. & Streckeisen, G. The leafspring seismometer: design and performance. *Bull. Seismol. Soc. Am.* **72**, 2349–2367 (1982).
12. Kamal & Mansinha, L. A test of the superconducting gravimeter as a long-period seismometer. *Phys. Earth Planet. Inter.* **71**, 52–60 (1992).
13. Smith, S. W. IRIS: A program for the next decade. *Eos* **67**, 213–219 (1986).
14. Dziewonski, A. M. & Woodhouse, J. H. An experiment in the systematic study of global seismicity: centroid-moment tensor solutions for 201 moderate and large earthquakes of 1981. *J. Geophys. Res.* **88**, 3247–3271 (1983).
15. Bullen, K. E. & Bolt, B. A. *An Introduction to the Theory of Seismology* 4th edn (Cambridge Univ. Press, 1985).
16. Beroza, G. & Jordan, T. Searching for slow and silent earthquakes using free oscillations. *J. Geophys. Res.* **95**, 2485–2510 (1990).
17. Turcotte, D. L. & Schubert, G. *Geodynamics Applications of Continuum Physics to Geological Problems* (Wiley, New York, 1982).

Acknowledgements. We thank R. Geller, S. Ida, M. Fujimoto and Y. Fukao for comments on this manuscript.

Correspondence and requests for materials should be addressed to N.K. (e-mail: shibata@geo.titech.ac.jp).

Global and local measures of the intrinsic Josephson coupling in $\text{Ti}_2\text{Ba}_2\text{CuO}_6$ as a test of the interlayer tunnelling model

A. A. Tsvetkov[†], D. van der Marel^{*}, K. A. Moler[‡], J. R. Kirtley[§], J. L. de Boer^{*}, A. Meetsma^{*}, Z. F. Ren^{||}, N. Kolesnikov[¶], D. Dulic^{*}, A. Damascelli^{*}, M. Gruninger^{*}, J. Schützmann^{*}, J. W. van der Eb^{*}, H. S. Somal^{*} & J. H. Wang^{||}

^{*} Materials Science Centre, University of Groningen, 9747 AG Groningen, The Netherlands

[†] P. N. Lebedev Physical Institute, 117924 Moscow, Russia

[‡] Department of Physics, Princeton University, Princeton, New Jersey 08544, USA

[§] IBM T. J. Watson Research Center, PO Box 218, Yorktown Heights, New York 10598, USA

^{||} Department of Chemistry, SUNY at Buffalo, Buffalo, New York 14260-3000, USA

[¶] Institute of Solid State Physics, Russian Academy of Sciences, Chernogolovka, Russia

One leading candidate theory of high-temperature superconductivity in the copper oxide systems is the interlayer tunnelling (ILT) mechanism¹. In this model, superconductivity is created by tunnelling of electron pairs between the copper oxide planes—contrasting with other models in which superconductivity first arises by electron pairing within each plane. The ILT model predicts that the superconducting condensation energy is approximately equal to the gain in kinetic energy of the electron pairs due to tunnelling. Both these energies can be determined independently^{2–4}, providing a quantitative test of the model. The gain in kinetic energy of the electron pairs is related to the interlayer plasma frequency, ω_p , of electron pair oscillations, which can be measured using infrared spectroscopy. Direct imaging of magnetic flux vortices also provides a test⁵, which is performed here on the same samples. In the high-temperature superconductor $\text{Ti}_2\text{Ba}_2\text{CuO}_6$, both the sample-averaging optical probe and the local vortex imaging give a consistent value of $\omega_p \approx 28 \text{ cm}^{-1}$ which, when combined with the condensation

energy, produces a discrepancy of at least an order of magnitude with deductions based on the ILT model.

In the ILT model, the normal state is different in nature from the traditional Landau Fermi liquid. As a result, coherent transport of single charge carriers between the planes is strongly inhibited in the normal state. In the superconducting phase, tunnelling of pairs is possible, and the superconducting condensation energy (E_{cond}) in the ILT model is the gain in kinetic energy (E_I) due to the tunnelling of those pairs: $E_{\text{cond}} = \eta E_I$. The number η is of the order of 1 when ILT is the only active pairing mechanism³. With conventional mechanisms, although usually $\eta \ll 1$, there is no prediction for η that is free from materials parameters. A crucial point in this discussion is that both E_{cond} and E_I are experimentally accessible quantities, thus allowing the experimental verification of the ILT hypothesis. E_{cond} can be measured from the specific heat⁶; E_I can be determined by measuring the interlayer (Josephson) plasma frequency⁷.

For this work, we used two kinds of samples: single crystals and epitaxial thin films of $\text{Ti}_2\text{Ba}_2\text{CuO}_6$. The crystals have a superconducting transition temperature (T_c) of 82 K and transition width (10%–90%) of 13 K, as determined by bulk susceptibility measurements using a superconducting quantum interference device (SQUID). Using 4-circle X-ray diffraction, we verified that the material belongs to the tetragonal $I4/mmm$ space group, with (for the crystals) lattice parameters $a = b = 3.867 \text{ Å}$, and $c = 23.223 \text{ Å}$. The films have $T_c = 80 \text{ K}$ as determined by d.c. resistivity, and $c = 23.14 \text{ Å}$. Both types of samples have relatively large physical dimensions perpendicular to the c -axis, corresponding to the conducting copper oxide planes (50 mm^2 for the thin films, and 1 mm^2 for the crystals). They have small dimensions along the c -axis ($\sim 1 \text{ μm}$ for the thin films, and $\sim 50 \text{ μm}$ for the crystals).

To determine the plasma resonance, we measure the reflection coefficient of infrared radiation incident on the a – b plane at a large angle (80°) with the surface normal⁸. A sketch of our experiments is shown in Fig. 1. In the case of the single crystals, the intensity of the reflected light drops below our detection limit if the wavelength exceeds 0.2 mm (that is, for $\omega/2\pi c < 50 \text{ cm}^{-1}$, where ω is an angular frequency and c is the speed of light) due to diffraction. Using thin films we were able to extend this range to 20 cm^{-1} . The electric field vector of the radiation is chosen to be parallel to the plane of incidence, resulting in a large component perpendicular to the CuO_2 planes. This geometry allows absorption of the light by lattice vibrations and plasma-oscillations polarized perpendicular to the planes.

In Fig. 2a we show the single-crystal and thin-film reflectivity for $\omega/2\pi c$ above and below 150 cm^{-1} , respectively. All prominent absorption lines for frequencies above 50 cm^{-1} correspond to infrared-active lattice vibrations, which show no strong temperature dependence. In the 4 K spectrum we observe a clear absorption

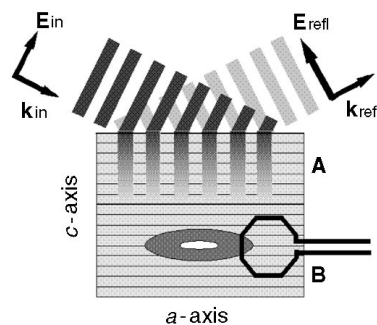


Figure 1 Diagram of the experimental set-up. **A**, Grazing-incidence reflectivity. The p-polarized light incident at a grazing angle sets up a periodic electric field pattern, which is polarized perpendicular to the sample surface, and decays exponentially inside the solid. **B**, Scanning SQUID microscopy. The octagonal pick-up loop detects the magnetic flux perpendicular to the a – c face.

at $\omega_J = 27.8 \text{ cm}^{-1}$. This resonance exhibits a strong red shift on raising the temperature, as shown in Fig. 2b. Above 70 K, it has shifted outside our spectral window. In Fig. 2c we also present the temperature dependence of the resonance position, $\omega_J(T)$, squared and normalized to 4 K, $\omega_J(T)^2/\omega_J(4 \text{ K})^2$. This temperature dependence extrapolates to zero at T_c , which indicates that it is a plasma resonance of the paired charge carriers. We therefore attribute this absorption to a Josephson plasmon, a collective oscillation of the paired charge carriers perpendicular to the coupled superconducting planes⁷. For a purely electronic system the Josephson resonance frequency, c/λ_c , where λ_c is the c -axis penetration depth, is determined by the supercurrent density along the c -axis. Because in the present case the Josephson plasma resonance is located at a frequency below the infrared-active lattice vibrations, the corresponding dynamical electric field is screened by the ions and the lattice vibrations, characterized by a dielectric constant ϵ_{cs} . As a result we observe the Josephson resonance at a reduced frequency $\omega_J = \epsilon_{cs}^{-1/2} c/\lambda_c$. We performed a full optical analysis of these spectra in the spectral range 20–6,000 cm^{-1} using Fresnel's equations for reflection at an oblique angle of incidence of anisotropic optical media. This way we were able to extract the dielectric function ϵ_{cs} from our data. For frequencies below 40 cm^{-1} , $\epsilon_{cs} = 11.3 \pm 0.5$. We therefore obtain $\lambda_c(4 \text{ K}) = 17.0 \pm 0.3 \text{ }\mu\text{m}$.

An independent experimental measure of the interlayer coupling is provided by a direct measurement of λ_c . Here we employ the fact that vortices which are orientated parallel to the planes, called interlayer Josephson vortices, have a characteristic size λ_c along the

planes and λ_a perpendicular to the planes⁹. In order to determine λ_c directly, we used a scanning SQUID microscope¹⁰ to map the magnetic fields perpendicular to an a - c face at 4 K (Fig. 1B)⁵. The crystal was cooled in a magnetically shielded cryostat with a residual magnetic field of a few milligauss, resulting in the presence of a few isolated trapped vortices (Fig. 3a). The jitter apparent in this image is due to the mechanical scanning mechanism used in our SQUID microscope. With an octagonal pick-up loop of diameter $L = 4 \text{ }\mu\text{m}$, the vortices were resolution-limited along the c direction ($\lambda_c \ll L$), but not along the a direction ($\lambda_c > L$).

Fitting the longitudinal cross-sections (Fig. 3b) to the functional form for the magnetic fields of an interlayer Josephson vortex⁹ convoluted with the shape of the pick-up loop^{5,10} gave the results $\lambda_c = 17 \pm 4 \text{ }\mu\text{m}$ and $\lambda_c = 19 \pm 1 \text{ }\mu\text{m}$ for these two vortices. The statistical error bars were determined using a criterion of doubling of the variance from the least-squares value, but systematic errors from the background and the shape of the pick-up loop, and the effect of the surface on the shape of the vortex, may be as large as 30%. Using this technique, we imaged vortices in three pieces cut from a large single crystal, which was part of the mosaic used to make the measurements in Fig. 1A. The vortices in all three pieces confirm the plasma resonance frequency of $\sim 28 \text{ cm}^{-1}$. This agreement rules out alternative interpretations of our results¹¹.

We are now ready to determine E_J using the relation⁷ between ω_J and E_J $\omega_J^2 \hbar^2 = 4\pi d a^{-2} e^2 E_J$, where d is the distance between planes (11.6 Å), a is the cell parameter (3.87 Å), $\hbar$ is the Planck constant, and $e^* = 2e$ is the charge of the pairs. The result is

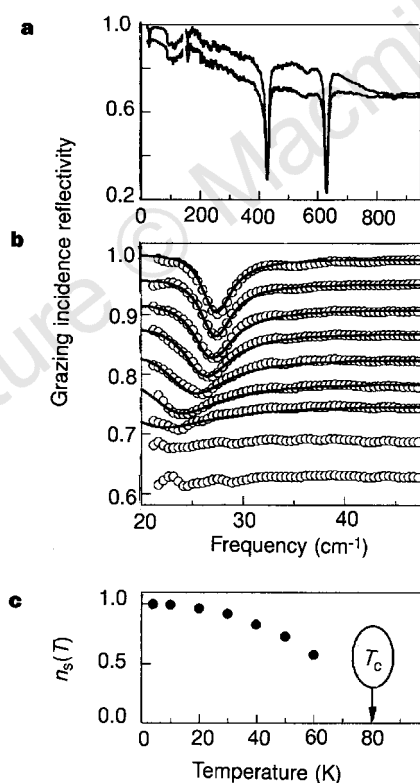


Figure 2 Optical measurements. **a** P-polarized reflectivity at 80° angle of incidence of $\text{Ti}_2\text{Ba}_2\text{CuO}_6$ ($T_c = 82 \text{ K}$) at 4 K (upper curve) and 100 K (lower curve). Frequencies above (below) 150 cm^{-1} correspond to single-crystal (thin-film) data. **b**, Thin-film spectra on an expanded frequency scale. From top to bottom: 4 K, 10 K, 20 K, 30 K, 40 K, 50 K, 60 K, 75 K and 90 K. The curves have been given incremental 3% vertical offsets for clarity. The solid curves correspond to calculations as described in the text. **c**, Temperature dependence of the superfluid density $n_s(T) = \omega_J(T)^2/\omega_J(4 \text{ K})^2$, demonstrating that the resonance frequency converges to zero at T_c .

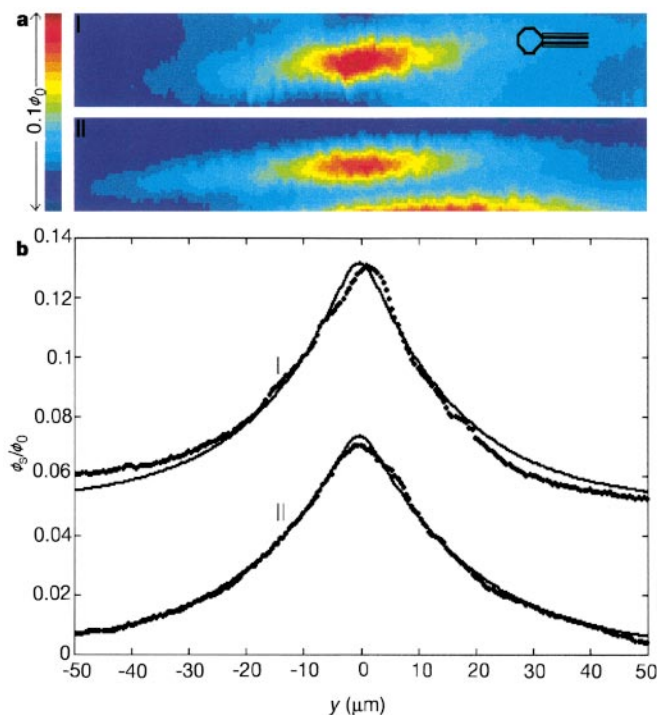


Figure 3 Local magnetic measurements with a scanning SQUID. **a**, Images of the magnetic field perpendicular to an a - c face of a $\text{Ti}_2\text{Ba}_2\text{CuO}_6$ single crystal in two different locations, showing two different interlayer Josephson vortices. The dashed lines indicate the longitudinal cross-sections. Inset: sketch of the 4- μm octagonal pick-up loop. **b**, The flux through the SQUID pick-up loop along the longitudinal cross-sections. ϕ_s is the magnetic flux through the SQUID pick-up loop; $\phi_0 = 20.7 \text{ G }\mu\text{m}^2$ is the superconducting flux quantum. The solid curves are fits which determine the c -axis penetration depths of these two vortices to be (I) $\lambda_c = 17 \pm 4 \text{ }\mu\text{m}$ and (II) $\lambda_c = 19 \pm 1 \text{ }\mu\text{m}$.

$E_j = 0.24 \mu\text{eV}$ per formula unit. For $\text{Ti}_2\text{Ba}_2\text{CuO}_6$, the measured value of E_{cond} is $100 \pm 20 \mu\text{eV}$ per formula unit⁶. Hence $\eta = E_j/E_{\text{cond}} = 0.0024 \pm 0.0005$, which is clearly at variance with the notion that condensation in the high- T_c superconductors is due to the gain in kinetic energy of the pairs (E_j) in the superconducting state. □

Received 13 March; accepted 29 June 1998.

1. Anderson, P. W. *The Theory of Superconductivity in the High- T_c Cuprates* (Princeton Univ. Press, 1997).
2. Anderson, P. W. Interlayer tunnelling mechanism for high- T_c superconductivity: comparison with c-axis infrared experiments. *Science* **268**, 1154–1155 (1995).
3. Leggett, A. J. Interlayer tunnelling models: implications of a recent experiment. *Science* **274**, 587–590 (1996).
4. van der Marel, D., Schützmann, J., Somal, H. S. & van der Eb, J. W. in *Proc. 10th Anniversary HTS Workshop on Physics, Materials & Applications* (eds Batlogg, B., Chu, C. W., Chu, W. K., Gubser, D. U. & Müller, K. A.) 357–370 (World Scientific, Rivers Edge, NJ, 1996).
5. Moler, K. A., Kirtley, J. R., Hinks, D. G., Li, T. W. & Xu, M. Images of interlayer Josephson vortices in $\text{Ti}_2\text{Ba}_2\text{CuO}_{6+\delta}$. *Science* **279**, 1193–1195 (1998).
6. Loram, J., Mirza, K. A., Wade, J. M., Cooper, J. R. & Liang, W. Y. The electronic specific heat of cuprate superconductors. *Physica C* **235**, 134–137 (1994).
7. Lawrence, W. E. & Doniach, S. in *Proc. 12th Int. Conf. on Low Temperature Physics* (ed. Kanda, E.) 361–362 (Academy, Kyoto, 1971).
8. Schützmann, J. *et al.* Experimental test of the inter-layer pairing models for high- T_c superconductivity using grazing incidence infrared reflectometry. *Phys. Rev. B* **55**, 11118–11121 (1997).
9. Clem, J. R. & Coffey, M. W. Viscous flux motion in a Josephson-coupled layer model of high- T_c superconductors. *Phys. Rev. B* **42**, 6209–6216 (1990).
10. Kirtley, J. R. *et al.* High-resolution scanning SQUID microscope. *Appl. Phys. Lett.* **66**, 1138–1140 (1995).
11. Anderson, P. W. c-Axis electrodynamics as evidence for the interlayer theory of high-temperature superconductivity. *Science* **279**, 1196–1197 (1998).

Acknowledgements. We thank M. Bhushan and M. Ketchen for assistance with the development of the scanning SQUID microscope; P. W. Anderson for comments; and P. Stamp for reading the manuscript. K.A.M. acknowledges the support of an R. H. Dicke postdoctoral fellowship. This work was supported by the Netherlands Foundation for Fundamental Research on Matter (FOM) with financial aid from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). A.A.T. acknowledges the support of NWO, RFBR and the Russian Superconductivity Program. The work performed at SUNY-Buffalo was partly supported by NYSERDA.

Correspondence and requests for materials should be addressed to D.v.d.M. (e-mail: D.van.der.marel@phys.rug.nl).

Ultra-low-threshold field emission from conjugated polymers

I. Musa, D. A. I. Munindrasdasa, G. A. J. Amaratunga & W. Eccleston

Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool L693BX, UK

Field-emission displays contain materials that emit electrons when charged to a low (negative) potential; the electrons excite light emission from phosphor screens. These devices have the potential to provide flat-panel visual displays with good picture quality at low power consumption and low cost¹. Field-emission devices at present use arrays of microfabricated tips as the emitting cathodes, but a potentially cheaper and simpler alternative is to use a thin-film cathode. This requires the identification of materials that will emit an appreciable electron current at low applied fields. Nitrogen-doped, chemical-vapour-deposited diamond films² and amorphous carbon films³ have been explored for this purpose. The low electron affinity⁴, wide bandgap and excellent transport properties⁵ of some conducting organic polymers suggest that they might also provide good cathode materials. Here we demonstrate that this is so, reporting field emission from thin films of regioregular poly(3-octylthiophene) deposited on n-doped silicon, with indium tin oxide as the anode. The threshold fields that we measure for electron emission from these films are the lowest yet reported for any carbon-based material.

Regioregular poly(3-octylthiophene) (P3OT) pellets were synthesized by the technique reported by Chen *et al.*⁶. Dissolving 5 mg of P3OT in 1 ml of chloroform formed films of thickness 5 μm .

The solution was then cast onto a pre-cleaned $1 \text{ cm} \times 1 \text{ cm}$ highly doped n-Si substrates (resistivity $\approx 0.003\text{--}0.01 \Omega \text{ cm}$) in a clean non-vacuum atmosphere. The optical Tauc bandgap was determined, at room temperature, to be $\sim 1.75 \text{ eV}$ by ultraviolet-visible spectrophotometry (this bandgap is defined as the intercept of the gradient of a $(\alpha E)^{1/2}$ versus E plot with the E axis, where α is absorption coefficient and E the photon energy). This estimate agrees well with those reported elsewhere⁶. In addition, a single broad absorption maximum was observed at $\sim 2.38 \text{ eV}$, which suggests that these 'as synthesized' films are only slightly doped. Nuclear magnetic resonance measurements indicate a regular 'head-tail' arrangement of the alkyl side chains (which can be thought of as a measure of how ordered the polymers are) of $>90\%$. Following casting the P3OT/n⁺⁺-Si devices were immediately placed in a vacuum of pressure $\sim 10^{-7}$ torr to dry for 24 hours.

The field-emission experiments were carried out in the flat plate configuration with an indium tin oxide coated glass anode. The separation between the anode and cathode was varied between 27 μm and 130 μm using glass-fibre spacers. The pressure in the vacuum was $\sim 2 \times 10^{-7}$ torr. Figure 1 shows typical current-voltage (I - V) characteristics. Two regions are clearly identified (called here 1 and 2). Region 1a and 1b denotes the emission before 'conditioning', and region 2 is after 'conditioning'. A highly stable and reproducible current level was maintained thereafter. Here conditioning refers to a little-understood phenomenon related to initial electric-field cycling required for cathodes before they show stable emission. It is thought to be related to field-induced removal of surface adsorbates. In the case of polymers, we also speculate that there are possible chemical excitations taking place at the emission sites during conditioning (see later). In our experiments the conditioning field was always below $10 \text{ V } \mu\text{m}^{-1}$ in order to avoid any influence of vacuum breakdown phenomena, which can occur at fields above $15 \text{ V } \mu\text{m}^{-1}$ at the pressure used.

The remarkable current density (J)-electric field (F) characteristics, with threshold field of only $\sim 0.2 \text{ V } \mu\text{m}^{-1}$ for a current density of $1 \mu\text{A cm}^{-2}$, are shown in Fig. 2a. Figure 2b shows the J - F characteristics as a function of anode-cathode spacing. In addition, a slope of 1.8 of the log I -log V plot indicates that a space-charge-limited current (SCLC) may be limiting the emission process. The measured anode voltage versus gap separation is shown in Fig. 2a inset and indicates a nonlinear relationship. This shows that surface and bulk charge properties may be important to the emission process.

The surface morphology of the film was examined by scanning electron microscopy (SEM) both before and after field emission (Fig. 3a-d). There was no apparent explosive destruction to the film which can be associated with discharge current phenomenon during field emission. We also examined the stability of the emission current over time. The emission characteristic recorded over a 16-hour period of continuous emission is shown in Fig. 4. The initial current drops to 55% of its value in 2 hours, and thereafter remains stable. The sensitivity of the emission to a change in pressure was also examined: the emission current dropped to below 10 pA when the pressure rose to 10^{-3} torr.

The polymer films show void-like features on the surface, of density $\sim 4 \times 10^8 \text{ cm}^{-2}$ and of sizes ranging from a $0.40 \mu\text{m}$ to few micrometres. These could be sites for preferential emission due to electric field intensification at the edges of the voids. Normally, surface field emission is associated with a tunnelling process at the surface, first proposed by Fowler and Nordheim⁷. The relationship between the current density and electric field will then be given by $J/(\beta V)^2 \propto \exp(-\kappa\phi/\beta V)$, where κ is a constant, ϕ is the barrier height and β is a factor (m^{-1}) which converts the local voltage V to the local electric field (βV); a value of $\beta = 1/d$ where d is the anode-cathode separation, signifies an ideal flat surface.

It is usual to plot $\ln(J/F^2)$ as a function of $1/F$. This type of plot is shown in Fig. 1 inset for region 1b, before conditioning, where a

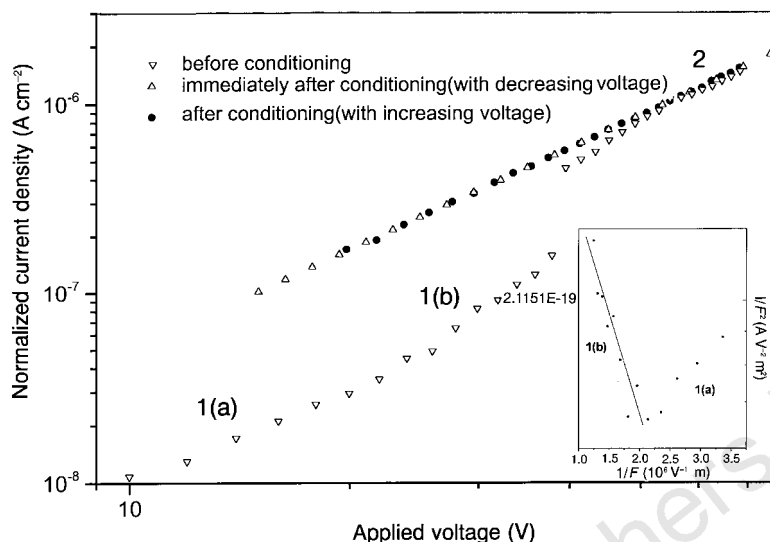


Figure 1 Field emission current density versus voltage. A logarithmic plot of normalized current density as a function of applied voltage for a typical 'as synthesized' regioregular poly(3-octylthiophene) 5- μm -thick film with a gap spacing of 47 μm . The current density is calculated by dividing the total current collected by the plate anode by the area of the cathode it covers, typically 1 cm^2 .

Two regions are clearly observed; region **1** and **2**. Inset, Fowler-Nordheim plot (J/F^2 versus $1/F$) for region **1**. An approximately linear relationship is only seen for region **1b**. With an enhancement factor of ~ 50 , a barrier of 0.1 eV is estimated, as described in the text. Region **2** follows conditioning; there is no hysteresis and the measurements are highly reproducible.

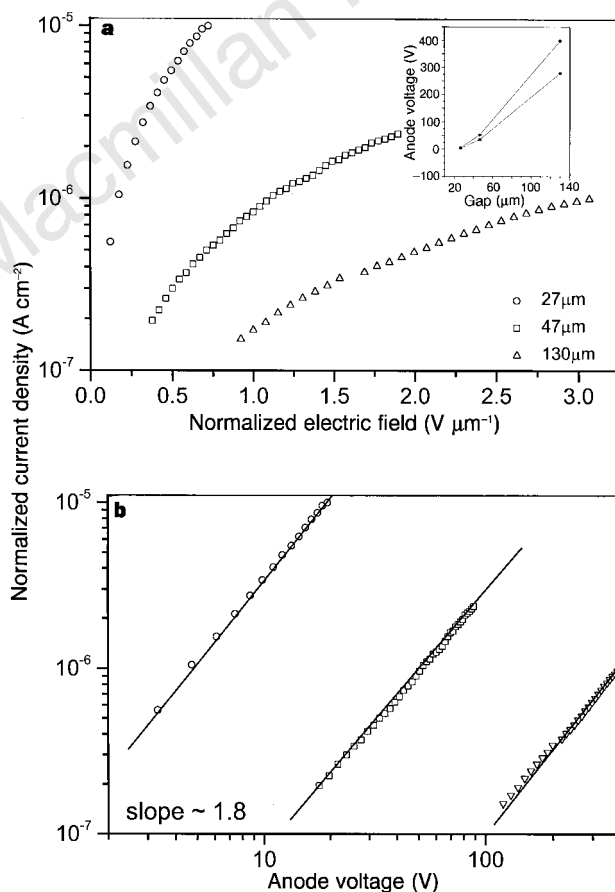


Figure 2 The effect of varying the anode-cathode distance. **a**, Comparison of emission for fibre spacing of 27 μm , 47 μm and 130 μm . The threshold field is as low as 0.2 $\text{V } \mu\text{m}^{-1}$ for the 27- μm spacing. Normalized current density is defined by assuming that the emission occurred continuously over the entire film (1 $\text{cm} \times 1 \text{ cm}$), although this will give the minimum possible current density as there is probably preferential emission at sites, as explained in the text. In addition, the plot indicates that a substantial portion of voltage must be across the film because the normalized current density versus normalized electric field

curves do not coincide with each other. The normalized electric field is defined as the anode voltage divided by the fibre spacing. The inset illustrates the nonlinear behaviour of the anode voltage with gap separation for two current values (upper trace, 1×10^{-6} A; lower trace, 5×10^{-7} A). This suggests that the polymer film cannot be treated as an ideal dielectric, because (from Gauss's law) a linear relationship is expected. **b**, I - V characteristics for different gap spacing. A slope of ~ 1.8 is obtained. Space-charge limited current may be a possible limiting mechanism in the field emission process (see text).

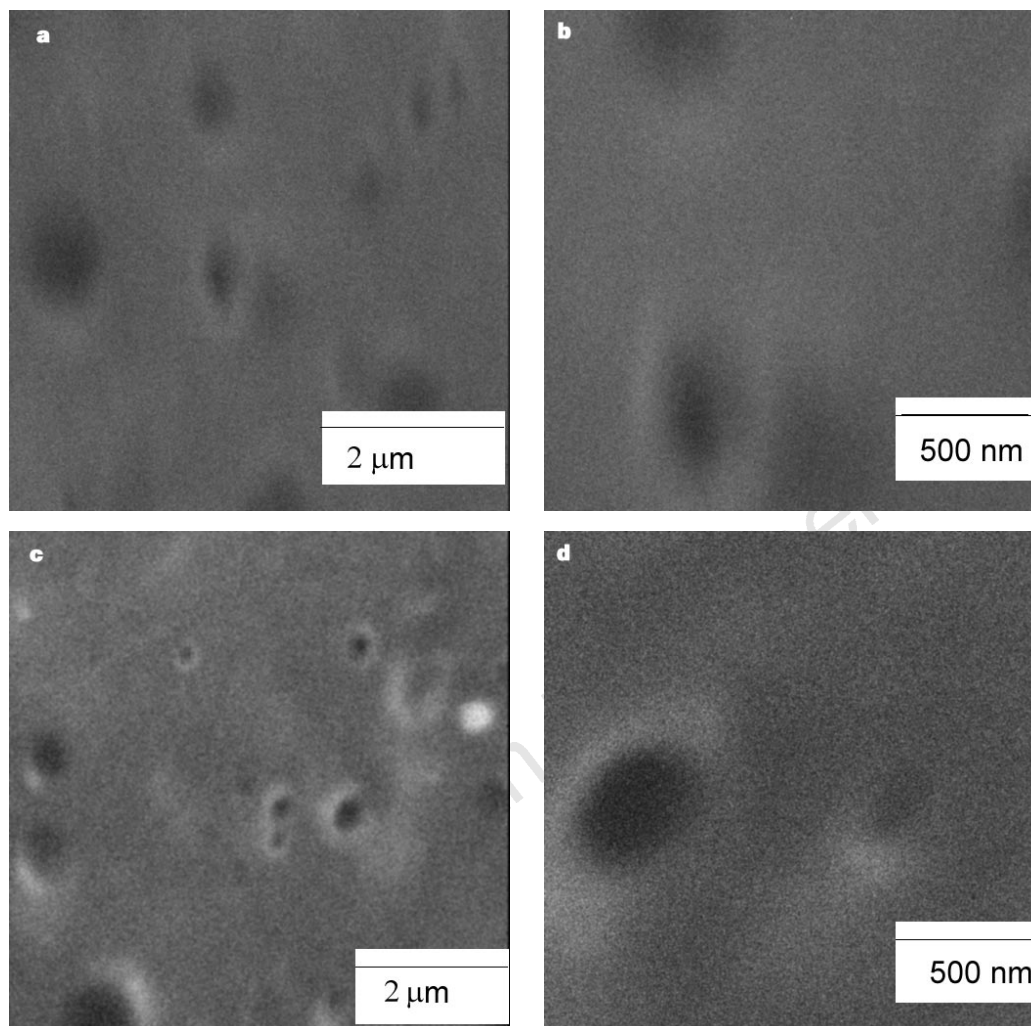


Figure 3 The morphology of the polymer surface both before (a, b) and after (c, d) field emission.

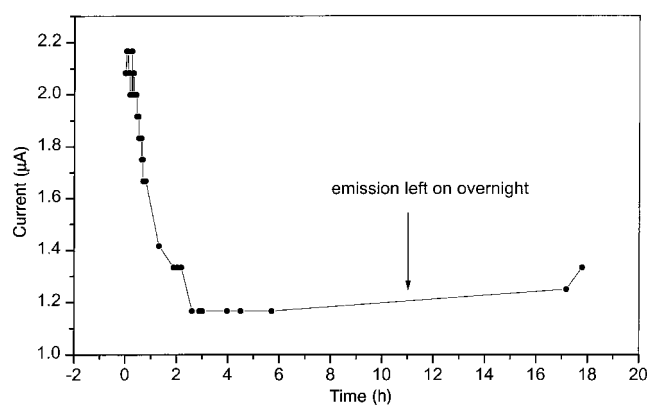


Figure 4 Life-time study of continuous field emission over a period of 16 hours. The current drops to 55% of its initial value in 2 hours, and remains stable thereafter.

straight line is observed at higher fields. If the conditions for emission from an ideal flat surface apply, with the voltage drop entirely across the vacuum gap, the barrier is estimated to be ~ 0.002 eV. This is an unrealistically low value. Assuming that emission occurs at the void edges and that there is an intensification factor $\beta \sim 50/d$ —based on an edge curvature of typically 20 nm and an edge depth of 1 μ m, estimated from SEM measurements—the

barrier is 0.1 eV. It therefore appears that the emission process is not simply a Fowler–Nordheim tunnelling mechanism at the surface from an ideal thin film, and that other contributing factors are involved. With materials of relatively small positive electron affinity, there can be substantial potential field drop across the emitting material resulting in band bending in such a way that there are two barriers; one at the back surface of the material and one at the

vacuum (emitting) surface. Electrons entering the film from the back contact can be heated (that is, enter energy states far from the conduction band edge) during transport to the vacuum (emitting) surface due to the high field within the film. The probability of electron emission will be high at the surface if such electron heating occurs, because the electrons are energetically closer to the vacuum level. This may explain the apparently low barrier obtained for the Fowler–Nordheim region in Fig. 1. For field emission from polymers, it appears that there is a small positive barrier—according to the Fowler–Nordheim relationship—before ‘conditioning’. But after conditioning, the J – F data, when plotted in the Fowler–Nordheim axes, has no region indicating a negative slope. This indicates that field emission is no longer limited by a positive barrier at the polymer–vacuum interface. Possible limiting mechanisms are the flow of SCLC in the bulk of the film, or in the vacuum immediately above the emission spot.

The gradient of 1.8 obtained from the log I –log V plot of Fig. 2 does not make it clear whether one type of space-charge-limiting mechanism dominates over the other. A slope of 1.5 is predicted by the Langmuir–Child⁸ law for SCLC in a vacuum, and a slope of at least 2.0 is predicted for SCLC in the film bulk from the Lampert–Rose model⁹. We note that a very similar gradient value of ~ 1.75 has been measured for nitrogenated diamond³, which showed the lowest emission threshold field of any carbon-based material before the present results from conjugated polymers. It is known that C–H bonding in polymeric carbon materials can lead to negative electron affinity, as in the case of H-terminated diamond {111} and {100} surfaces¹⁰. The relative energies of the conduction band edge and the vacuum level are largely determined by C–H bonding at the surface. We speculate that the conditioning in Fig. 1 reflects a change in polymer structure at the high-field sites from which electron emission originates. The concentration of carriers and electric field at these local emission sites could in principle lead to a significant rise in the local temperature. If the transformation of the polymer gives it a negative or very low electron affinity at the emission sites, then post-conditioning current characteristics similar to those of diamond would be expected¹¹.

The extremely low threshold field for electron emission in conjugated polymers offers the prospect of an all-polymer flat-panel display, in which polymer thin-film transistors are used for switching the pixels, and polymer field-emission cathodes stimulate light emission from phosphor screens.

Note added in proof: Since submission of this Letter, an electron emission process based on a triple junction effect at a diamond/metal/vacuum interface has been proposed¹¹. Such a process may also be relevant for emission from the surface of the polymer void. □

Received 14 May; accepted 15 June 1998.

- Chalamala, B. R., Wei, Y. & Gnade, B. E. FED up with fat tubes (review of current technology). *IEEE Spectrum* **35**, 42–51 (1998).
- Okano, K., Koizumi, S., Silva, S. P. & Amaratunga, G. A. J. Low-threshold cold cathodes made of nitrogen-doped chemical-vapour-deposited diamond. *Nature* **381**, 140–141 (1996).
- Amaratunga, G. A. J. & Silva, S. R. P. Nitrogen containing hydrogenated amorphous carbon for thin-film field emission cathodes. *Appl. Phys. Lett.* **68**, 2529–2531 (1996).
- Bredas, J. L. Relationship between band-gap and bond length alternation in organic conjugated polymers. *J. Chem. Phys.* **82**, 3808–3811 (1985).
- Nishino, H., Yu, G., Heeger, A. J., Chen, T.-A. & Rieke, R. D. Electroluminescence from blend films of poly(3-hexylthiophene) and poly(N-vinylcarbazole). *Synth. Met.* **68**, 243–247 (1995).
- Chen, T.-A., Wu, X.-M. & Rieke, R. D. Regiocontrolled synthesis of poly(3-alkylthiophenes) mediated by Rieke zinc: their characterisation and solid state properties. *J. Am. Chem. Soc.* **117**, 233–244 (1995).
- Fowler, R. H. & Nordheim, L. Electron emission in intense fields. *Proc. R. Soc. Lond. A* **119**, 173–181 (1928).
- Anderson, W. A. Role of space charge in field emission cathodes. *J. Vac. Sci. Technol.* **11**, 383–388 (1993).
- Lampert, M. A. & Mark, P. *Current Injection in Solids* 44–46 (Academic, New York, 1970).
- Robertson, J. Band model for electron emission from diamond and diamond-like carbon. *Symp. Proc. MRS* **422**, 777–782 (1996).
- Geis, M. W. *et al.* *Nature* **393**, 431–435 (1998).

Acknowledgements. We thank N. Rupasinghe for assistance during the measurements, M. Chhowalla for discussions, and K. Molloy for technical assistance. We also thank the UK Engineering and Physical Research Council for financial support through studentships.

Correspondence and requests for materials should be addressed to G.A.J.A. (e-mail: amaratun@liv.ac.uk).

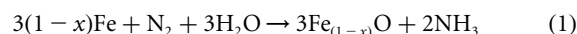
Abiotic nitrogen reduction on the early Earth

Jay A. Brandes, Nabil Z. Boctor, George D. Cody, Benjamin A. Cooper, Robert M. Hazen & Hatten S. Yoder Jr

Geophysical Laboratory, Carnegie Institution of Washington, 5251 Broad Branch Road NW, Washington DC 20015-1305, USA

The production of organic precursors to life depends critically on the form of the reactants. In particular, an environment dominated by N_2 is far less efficient in synthesizing nitrogen-bearing organics than a reducing environment rich in ammonia (refs 1, 2). Relatively reducing lithospheric conditions on the early Earth have been presumed to favour the generation of an ammonia-rich atmosphere, but this hypothesis has not been studied experimentally. Here we demonstrate mineral-catalysed reduction of N_2 , NO_2^- and NO_3^- to ammonia at temperatures between 300 and 800 °C and pressures of 0.1–0.4 GPa—conditions typical of crustal and oceanic hydrothermal systems. We also show that only N_2 is stable above 800 °C, thus precluding significant atmospheric ammonia formation during hot accretion. We conclude that mineral-catalysed N_2 reduction might have provided a significant source of ammonia to the Hadean ocean. These results also suggest that, whereas nitrogen in the Earth’s early atmosphere was present predominantly as N_2 , exchange with oceanic, hydrothermally derived ammonia could have provided a significant amount of the atmospheric ammonia necessary to resolve the early-faint-Sun paradox³.

To assess the extent to which N_2 might have been reduced to NH_3 via mineral-catalysed reduction, we have conducted a series of high-temperature, high-pressure experiments. If the early Earth’s lithosphere was more reduced than it is today with a higher percentage of iron in the FeO state⁴, then reaction of this reduced iron with H_2O might have provided the necessary H_2 gas for N_2 reduction⁵. Furthermore, the reduction of N_2 to NH_3 might have been catalysed by iron oxides as these are used as the starting material for catalysts in industrial nitrogen reduction⁶. The first reaction considered was that of metallic iron with N_2 and H_2O to form NH_3 and iron oxides:



This reaction has been postulated to occur on the early Earth between metallic iron or reduced iron oxides within the lithosphere and degassing N_2 from the Earth’s interior⁷. In the present experiments the ratio of H_2O to Fe in the sample capsules was varied to generate a range of conditions from highly reducing ($H_2O:Fe = 1$) to mildly reducing ($H_2O:Fe \sim 1$ to 1.5) and oxidizing with excess H_2O ($H_2O:Fe > 1.5$). Under the most reducing conditions, at 700 °C and 0.1 GPa, up to 17 mol% of N_2 was reduced to NH_3 , whereas under mildly reducing conditions 1–3% of N_2 was reduced (Fig. 1). Under the most oxidizing conditions with excess H_2O present, only trace amounts of N_2 were reduced to NH_3 . Interestingly, no NH_3 was produced at 900 °C, even under the most reducing conditions, indicating the inherent thermodynamic instability of NH_3 at this high temperature.

To examine the system with more oxidized catalysts, experiments were conducted in N_2 atmosphere in which magnetite (Fe_3O_4) with formic acid was used as a reducing agent. It has long been established that, in the presence of a strong reductant (H_2 or hydride donor), pure synthetic magnetite acts as a catalyst for N_2 reduction⁶. Notably, magnetite is a major form of iron oxide in the present-day crust⁸. The potential for this reaction was explored with natural magnetite under geochemically relevant conditions spanning a range of temperatures and pressures. Yields of NH_3 for the $Fe_3O_4/N_2/HCO_2H$ system were comparable to those obtained under

equivalent reducing conditions from the $\text{Fe}/\text{N}_2/\text{H}_2\text{O}$ system (Fig. 2). Although the temperature of maximum yield dropped to 500°C , the upper thermal limit of reduction remained at $\sim 900^\circ\text{C}$. Increasing the pressure at 500°C had only a small effect (from 0.53% at 0.1 GPa to 0.58% at 0.4 GPa) on NH_3 yields. Low pressures (0.01 GPa) resulted in a marked decrease in NH_3 yields to less than 10% of comparable high-pressure values in both systems (Figs 1 and 2). In a similar manner to the $\text{Fe}/\text{H}_2\text{O}$ system, N_2 reduction was strongly inhibited by excess water (Fig. 2). Most gas compositions in metamorphic systems have ratios of H_2O to total gas of 0.5 or greater⁹, although gases trapped in some mid-ocean ridge basalts have been shown to have very low ratios of H_2O to total gas¹⁰. Thus, N_2 reduction would be restricted to localized regions from which H_2O had been removed by reaction with reduced minerals to form H_2 and oxides^{5,7}. A plausible scenario for the production of NH_3 in the crust involves the initial production of H_2 from the reaction of H_2O with reduced minerals, followed by the reduction of N_2 on the mineral surfaces. The high temperatures and pressures required in these experiments (0.1 GPa, $300\text{--}800^\circ\text{C}$) are consistent with this hypothesis.

The fate of the more oxidized forms of nitrogen, NO_2^- and NO_3^- , presumed to be in the Hadean ocean¹¹ (generated by atmospheric reactions driven by electrical discharge or cometary shock¹²) was also investigated. These compounds are of interest because they represent an additional source of reactive nitrogen to the Hadean ocean, whereas hydrothermal systems have been suggested as a possible sink for oxidized nitrogen¹¹. Experiments were conducted in which NO_2^- and NO_3^- solutions were reacted with mineral catalysts at high temperatures and pressures. We find that NO_2^- and NO_3^- in contact with iron sulphides are quickly and efficiently reduced to NH_3 at higher temperatures (Table 1). Reactions at 500°C reached 80% conversion efficiency within the 15 minutes required to bring the pressure vessel up to constant temperature (Table 1). Pyrite was as efficient as pyrrhotite in catalysing the conversion of NO_3^- to NH_3 and no added reductant (H_2S or H_2) was required. Magnetite and basalt, however, did not catalyse the reaction to NH_3 as efficiently or as rapidly as iron sulphides. Both magnetite and basalt quickly (<15 min at 500°C) converted NO_3^- and NO_2^- to species that could not be detected as solutes in aqueous solution. After 24 hours, significant amounts of NH_3 were produced, although yields were not as great as those found with iron sulphides. We hypothesize that the pathway for this reaction¹³ proceeds through NO or NO_2 , not N_2 , on the basis of the difficulty of N_2 reduction in the presence of excess water. Reduction of NO_3^- does not appear to pass through NO_2^- (as it does in the biological pathway¹⁴), as yields of NH_3 from NO_3^- are higher than those from

NO_2^- . Lower temperatures decreased the amount of conversion of nitrogen oxides to NH_3 . The addition of reducing agents within hydrothermal systems, however, might improve the conversion efficiency at low temperatures¹⁵. Because iron sulphides are ubiquitous in hydrothermal systems, it is likely that both NO_2^- and NO_3^- carried into such systems are eventually converted into NH_3 . After conversion, the NH_3 is quite stable, even at very high temperatures (700°C). Only at $T > 800^\circ\text{C}$ are both nitrogen oxides and NH_3 destroyed. Studies of hydrothermal systems indicate that high temperatures do not exceed 700°C (ref. 16). Thus, hydrothermal systems represent regions of the highest NH_3 conversion rates and stability on the Earth.

Three important conclusions follow from this study. First, the instability of NH_3 at $T > 800^\circ\text{C}$ strongly indicates that planetary bodies that undergo a 'hot accretionary' mechanism (including the Earth¹⁷) have their earliest atmospheres dominated by nitrogen in the N_2 form. Regardless of the degree to which the interior of the planet is reducing in nature, during the initial period of accretion where temperatures are above 800°C no NH_3 will form and any NH_3 derived from the thermal decomposition of organic matter in chondrites¹⁸ will be converted to N_2 . Indeed, most of the Archaean atmosphere is assumed to have formed under these conditions¹⁹. These results also indicate that surficial N_2 reduction would have been precluded by the high H_2O content and low pressures of the Hadean atmosphere. Thus it is likely that the very reducing conditions required for N_2 reduction would have been found only in high-pressure environments well beneath the Earth's surface.

Second, NH_3 produced by crustal N_2 reduction might have been an important source of reduced nitrogen to the Hadean ocean. After atmospheric formation a significant percentage of total Earth N is thought to have remained within the crust and mantle. Conversion of even a small percentage of this nitrogen to NH_3 after crustal temperatures dropped below 800°C would have provided a vast reservoir of reduced nitrogen for the early Earth. The total nitrogen budget of the Earth's crust and atmosphere is estimated to be 2×10^{20} mol (ref. 19). Conservatively assuming that only 25% of the total nitrogen remained within the crust and mantle after initial atmosphere formation, and that of this only 1% was converted to NH_3 , then it is calculated that a reservoir of 5×10^{17} mol N in reduced form would have resulted. In the absence of any losses or other inputs, this NH_3 was presumably transferred to the ocean from the crust by hydrothermal systems over timescales of 10^7 to 10^9 years. Such a source of ammonia would have been comparable in magnitude ($10^9\text{--}10^{11}$ mol N yr^{-1}) to the proposed oceanic reduction of NO_3^- by Fe^{2+} (ref. 11) and terrestrial TiO_2 photochemical catalysis of N_2 (ref. 20). Transfer of crustal ammonia to the ocean might

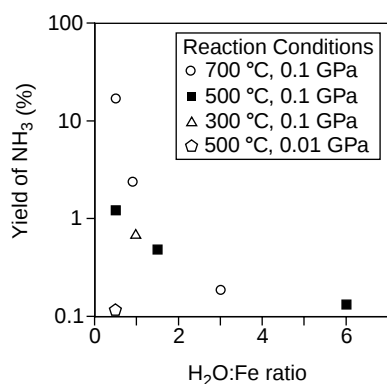


Figure 1 N_2 reduction to NH_3 plotted against initial $\text{Fe}:\text{H}_2\text{O}$ ratio at 300, 500 and 700°C . All reactions were conducted over a 24-h period. No detectable NH_3 was obtained in experiments conducted at 900°C or in experiments conducted at 0.01 GPa and $\text{H}_2\text{O}:\text{Fe}$ ratios larger than 0.5.

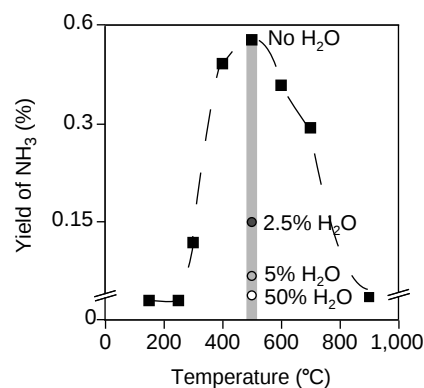


Figure 2 N_2 reduction to NH_3 plotted against temperature at 0.1 GPa in the $\text{HCO}_2\text{H}/\text{Fe}_3\text{O}_4$ system (squares). Also shown are yields at 500°C with different amounts of H_2O (given in mol%) added to the initial HCO_2H solution (circles). The 0.03% yield at 150°C represents the blank concentration of NH_3 . No detectable NH_3 above blank was obtained in experiments conducted at 0.01 GPa and 500°C .

Table 1 Conversion efficiencies to NH₃ from NO₂⁻ and NO₃⁻ solutions in contact with catalysts

Mineral	Compound	T (°C)	Time (h)	Solution type	Product yield (%)
Fe _(1-x) S	NO ₃ ⁻	300	24	H ₂ O	53
Fe _(1-x) S	NO ₂ ⁻	300	24	H ₂ O	21
Fe _(1-x) S	NO ₃ ⁻	500	24	H ₂ O	73
Fe _(1-x) S	NO ₂ ⁻	500	24	H ₂ O	37
Fe _(1-x) S	NO ₃ ⁻	700	24	H ₂ O	89
Fe _(1-x) S	NO ₃ ⁻	700	24	H ₂ O	74
Fe _(1-x) S	NO ₂ ⁻	900	24	H ₂ O	0
Fe _(1-x) S	NO ₃ ⁻	500	0.033	H ₂ O	74
FeS ₂	NO ₃ ⁻	500	24	H ₂ O	70
95%Fe, 5%Ni	NO ₃ ⁻	500	24	H ₂ O	41
95%Fe, 5%Ni	NO ₂ ⁻	500	24	H ₂ O	45
Fe ₃ O ₄	NO ₃ ⁻	900	24	H ₂ O	0
Fe ₃ O ₄	NO ₃ ⁻	500	24	H ₂ O	46
Fe ₃ O ₄	NO ₂ ⁻	500	24	H ₂ O	9.4
Fe ₃ O ₄	NO ₃ ⁻	500	0.033	H ₂ O	4.1
Fe ₃ O ₄	NO ₂ ⁻	500	0.033	H ₂ O	1.0
Basalt	NO ₃ ⁻	500	24	H ₂ O	20
Basalt	NO ₂ ⁻	500	24	H ₂ O	14
Fe _(1-x) S	NO ₃ ⁻	500	24	Sea water	84
Fe ₃ O ₄	NO ₃ ⁻	500	24	Sea water	1.2
Basalt	NO ₃ ⁻	500	24	Sea water	1.0

therefore have provided an early boost to oceanic NH₃ concentrations, and thus improved the prospects for synthesis and stability of N-containing organic compounds²¹. The results discussed here point to a probably oceanic or hydrothermal source for a large fraction of reduced nitrogen. Steady-state oceanic NH₃ concentrations would have depended on atmospheric loss rates^{3,11}; locally, however, hydrothermal environments and adjacent waters would clearly have exhibited the highest NH₃ concentrations in the prebiotic world. After the genesis of life, hydrothermal environments would have continued to be oases, providing NH₃ to early life forms until the advent of enzymatic systems capable of reducing NO₃⁻ and N₂ (ref. 22).

Last, these results on potential sources of ocean NH₃ must be factored into estimates of the NH₃ content of the Archaean atmosphere. Even low concentrations of atmospheric NH₃ could be of great importance to photolytically driven organic synthesis¹. Furthermore, Sagan and Chyba³ have demonstrated that an atmospheric mixing ratio of 10⁻⁵ for ammonia would have been sufficient, through greenhouse warming, to resolve the early-faint-Sun paradox. Maintenance of this mixing ratio over the first 10⁹ years of the Earth would have required a supply of 1.4 × 10¹⁸ g NH₃ (ref. 3). These results indicate that mineral-catalysed synthesis of NH₃ could have met at least one-third of this requirement, with the possibility that all of the necessary supply could have been produced in this manner if one assumes a highly reducing lithosphere during the Hadean. □

Methods

All experiments were run in sealed, N₂-purged, acid-washed, pure gold tubes. Iron sulphide and iron-metal catalysts were prepared from semiconductor-grade metals and pure elemental sulphur, with elemental compositions verified by ion-probe analysis. Magnetite catalyst was a very pure natural sample of fumarolic origin²³, whereas the basalt used was a natural alkali basalt (65992, ref. 28). Solution phases added to N₂ reduction capsules were either distilled deionized water or mixtures of formic acid and water. Solutions phases for NO₃⁻ and NO₂⁻ were 5 mM in either distilled deionized water or low-nutrient sea water. Tubes were loaded, purged with N₂ gas, welded shut and incubated in an internally heated pressure vessel under argon-medium^{24,25}. Tubes were opened, extracted with distilled water and analysed for NO₃⁻, NO₂⁻ and NH₃ by spectrophotometric methods^{26,27}.

Received 25 February; accepted 13 July 1998.

- Schlesinger, G. & Miller, S. L. Prebiotic synthesis in atmospheres containing CH₄, CO and CO₂. I. Amino acids. *J. Mol. Evol.* **19**, 376–383 (1983).
- Stribling, R. & Miller, S. L. Energy yields for hydrogen cyanide and formaldehyde syntheses: the HCN and amino acid concentrations in the primitive ocean. *Origins Life Evol. Biosph.* **17**, 261–273 (1987).

- Sagan, C. & Chyba, C. The early faint sun paradox: organic shielding of ultraviolet-labile greenhouse gases. *Science* **276**, 1217–1221 (1997).
- Kasting, J. F., Egger, D. H. & Raeburn, S. P. Mantle redox evolution and the oxidation state of the Archaean atmosphere. *J. Geol.* **101**, 245–257 (1993).
- Berndt, M. E., Allen, D. E. & Seyfried, W. E. Jr Reduction of CO₂ during serpentinization of olivine at 300 °C and 500 bar. *Geology* **24**, 351–354 (1996).
- Tamaru, K. in *Catalytic Ammonium Synthesis* (ed. Jennings, J. R.) 1–18 (Plenum, New York, 1991).
- Holland, H. D. *The Chemical Evolution of the Atmosphere and Oceans* (Princeton Univ. Press, 1984).
- Frost, B. R. Stability of oxide minerals in metamorphic rocks. *Rev. Mineral.* **25**, 469–483 (1991).
- Johnson, M. C., Anderson, A. T. Jr & Rutherford, M. J. Pre-eruptive volatile contents of magmas. *Rev. Mineral.* **30**, 281–330 (1994).
- Javoy, M. & Pineau, F. The volatiles record of a 'popping' rock from the mid-Atlantic ridge at 14° N: chemical and isotopic composition of gas trapped in the vesicles. *Earth Planet. Sci. Lett.* **107**, 598–611 (1991).
- Summers, D. P. & Chang, S. Prebiotic ammonia from reduction of nitrite by iron (II) on the early Earth. *Nature* **365**, 630–632 (1990).
- Mancinelli, R. L. & McKay, C. P. The evolution of nitrogen cycling. *Origins Life Evol. Biosph.* **18**, 311–326 (1990).
- Buresh, R. J. & Moraghan, J. T. Chemical reduction of nitrate by ferrous iron. *J. Environ. Qual.* **5**, 320–325 (1976).
- Knowlards, R. Denitrification. *Microbiol. Rev.* **46**, 43–70 (1982).
- Blöchl, E., Keller, M., Wachtershauser, G. & Stetter, K. O. Reactions depending on iron sulfide and linking geochemistry with biochemistry. *Proc. Natl Acad. Sci. USA* **89**, 8117–8120 (1992).
- Fournier, R. O. Conceptual models of brine evolution in magmatic hydrothermal systems. *US Geol. Surv. Prof. Pap.* **1350**, 1497–1506 (1987).
- Stevenson, D. J. in *Earth's Earliest Biosphere: Its Origin and Evolution* (ed. Schopf, J. W.) 32–40 (Princeton Univ. Press, 1983).
- Pizzarello, S., Feng, X., Epstein, S. & Cronin, J. R. Isotopic analyses of nitrogenous compounds from the Murchinson meteorite: ammonia, amines, amino acids, and polar hydrocarbons. *Geochim. Cosmochim. Acta* **58**, 5579–5587 (1994).
- Zhang, Y. & Zindler, A. Distribution and evolution of carbon and nitrogen in Earth. *Earth Planet. Sci. Lett.* **117**, 331–345 (1993).
- Henderson-Sellers, A. & Schwartz, A. W. Chemical evolution and ammonia in the early Earth's atmosphere. *Nature* **287**, 526–528 (1980).
- Bada, J. L. & Miller, S. L. Ammonium ion concentration in the primitive ocean. *Science* **159**, 423–425 (1968).
- Falkowski, P. G. & Raven, J. in *Aquatic Photosynthesis* Ch. 10 (Blackwell Science, Oxford, 1997).
- Zeis, E. G. The fumarolic incrustations in the valley of ten thousand smokers. *Nat. Geol. Soc. Cent. Tech. Papers, Katmai Series* **1**, no. 3 (National Geographic Society, Washington, 1924).
- Yoder, H. S. High–low quartz inversion up to 10,000 bars. *Trans. Am. Geophys. Union* **31**, 821–835 (1950).
- Kullerød, G. & Yoder, H. S. Pyrite stability in the Fe–S system. *Econ. Geol.* **54**, 534–550 (1959).
- Strickland, J. D. H. & Parsons, T. R. *A Practical Handbook of Seawater Analysis* (Fisheries Research Board of Canada, Ottawa, 1972).
- Jones, M. N. Nitrate reduction by shaking with cadmium. *Water Res.* **18**, 643–646 (1984).
- Yoder, H. S. & Tilley, C. E. Origin of basaltic magmas: an experimental study of natural and synthetic rock systems. *J. Petrol.* **3**, 342–532 (1962).

Acknowledgements. We thank M. L. Fogel, D. Rumble, C. Prewitt and C. Chyba for helpful discussion of this work. Benjamin Cooper, a summer intern at the Carnegie Institution, was killed in an car accident during the completion of research on this project; this paper is dedicated to his memory. This work was supported by NASA.

Correspondence and requests for materials should be addressed to J.A.B. (e-mail: brandes@gl.ciw.edu).

Quantification of dust-forced heating of the lower troposphere

P. Alpert*, Y. J. Kaufman†, Y. Shay-El*, D. Tanre‡, A. da Silva§, S. Schubert§ & J. H. Joseph*

* Department of Geophysics and Planetary Sciences, Tel-Aviv University, Israel 69978

† Climate and Radiation Branch, Code 913, NASA/GSFC, Greenbelt, Maryland 20771, USA

‡ Laboratoire d'Optique Atmosphérique, Bât. P5, UST de Lille, 59655-Villeneuve d'Ascq Cedex, France

§ Data Assimilation Office, Code 910.3, NASA/GSFC, Greenbelt, Maryland 20771, USA

Aerosols may affect climate through the absorption and scattering of solar radiation and, in the case of large dust particles, by interacting with thermal radiation^{1–3}. But whether atmospheric temperature responds significantly to such forcing has not been determined; feedback mechanisms could increase or decrease the effects of the aerosol forcing. Here we present an indirect measure of the tropospheric temperature response by explaining the 'errors' in the NASA/Goddard model/data-assimilation system. These errors, which provide information about physical processes

missing from the predictive model, have monthly mean patterns that bear a striking similarity to observed patterns of dust over the eastern tropical North Atlantic Ocean. This similarity, together with the high correlations between latitudinal location of inferred maximum atmospheric heating rates and that of the number of dusty days, suggests that dust aerosols are an important source of inaccuracies in numerical weather-prediction models in this region. For the average dust event, dust is estimated to heat the lower atmosphere (1.5–3.5 km altitude) by ~ 0.2 K per day. At about 30 dusty days per year, the presence of the dust leads to a regional heating rate of ~ 6 K per year.

It is widely held that cloud and water processes are predominant in prescribing climate model errors⁴. For instance, the ability of the general circulation models (GCMs) to describe faithfully the transport of water vapour into the upper troposphere by well developed clouds is a central point in the debate on the ability of GCMs to

predict the atmospheric warming due to the doubling of atmospheric CO_2 (ref. 5). On the other hand, most current atmospheric models do not include aerosol processes because aerosols are not properly monitored on a global scale⁶; their spatial distributions are therefore not known as required for their incorporation in operational GCMs^{7,8}. We now attempt to assess aerosol forcing by studying the errors of a GCM without aerosol physics within a data-assimilation system.

In principle, the best estimate of the state of the atmosphere at any given time (referred to here as the analysis) is obtained by data-assimilation systems in which the first guess, determined by a short term forecast with a state-of-the-art GCM, is optimally combined with all available observations⁹. The field differences between the analysis and the first guess are called the analysis increments (or updates) and result, in part, from a complex combination of all model errors due to inadequate representation of physical processes, that is, the model parameterizations, as well as from numerical errors and processes that have been ignored. These increments can be termed as 'model errors'. We note that in the implementation of the GEOS-1 assimilation scheme, the analysis increments are not inserted fully at each analysis time but instead they are gradually inserted over the six-hour period straddling the analysis time. This process is called incremental analysis update, or IAU. The IAU scheme avoids shocking the system, which can lead to spurious oscillations in the assimilating GCM¹⁰.

The study employs 5 years (1985–89) of the NASA GEOS-1 (ref. 11) analysis increments together with dusty days distributions over the Atlantic Ocean, as derived by Jankovik and Tanre¹². Dust is the most common aerosol type in the atmosphere. Persistent dust transport over the Atlantic Ocean has been measured in Barbados and Florida¹³. Dust dominates the NOAA images of aerosol over the ocean⁸. About half of the dust is estimated to be associated with anthropogenic sources³.

Figure 1 (right) shows the monthly averaged number of dusty days for 1984–88 for March, June, September and December, respectively, as derived by Jankovik and Tanre¹² using an aerosol optical retrieval method based on Meteosat data. It shows an annual cycle in which dust storms from the west Sahara propagate over the Atlantic region at about 11°N in March, gradually shifting to the north to 18°N in June (maximum northward location is at 21°N in August–September, Fig. 2). Also, there is a strong variation in intensity, as expressed by the number of dusty days.

For comparison, Fig. 1 (left) shows the corresponding incremental analysis updates for temperatures, IAU(T), integrated over the dust layer of 650–850 hPa. The 5-yr average is shifted here by 1 year to 1985–89 owing to data availability. The 650–850 hPa layer

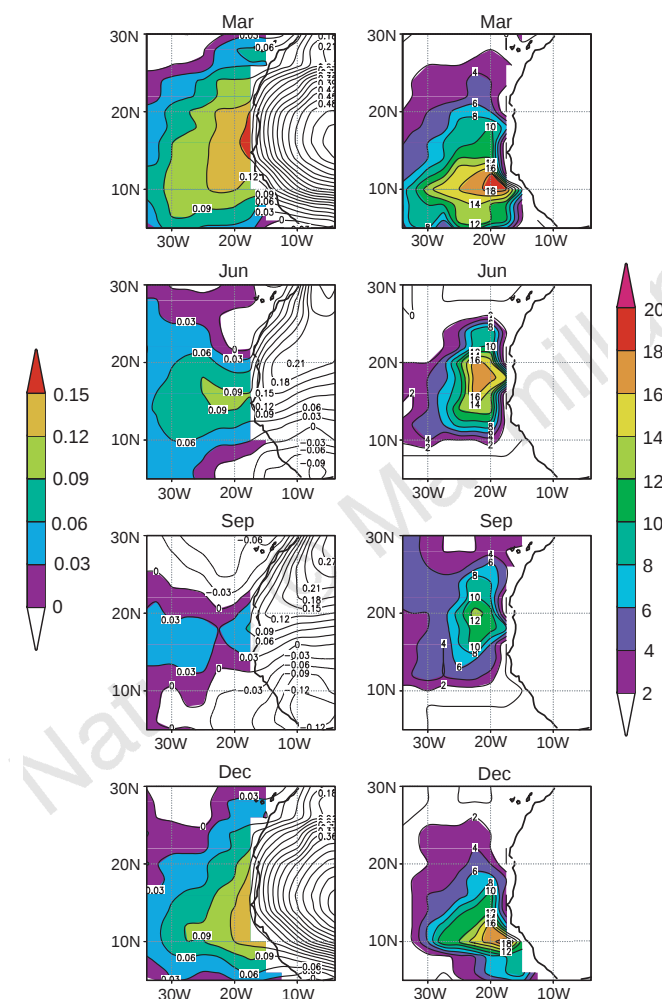


Figure 1 Right, monthly average image of the number of dusty days for five years from 1984 to 1988 for March, June, September and December, top to bottom; taken from ref. 12. Colour represents the number of days in steps of 2, as indicated on the bar to the right. Left, corresponding incremental analysis updates for temperatures, IAU(T), averaged during 1985–89, at the layer of 650–850 hPa. Shading corresponds to positive values over the ocean, which correspond to a warmer atmosphere than predicted by first guess from the model. Contour interval is 0.03 K d^{-1} . Shading calibration is shown on the colour bar on the left. The TIROS (Television Infra-Red Observational Satellite) Operational Vertical Sounder (TOVS) was the primary observational source for the IAU over the ocean²⁵. The low-tropospheric TOVS temperature estimates have a slight negative bias over the study domain²⁵, which suggests that the IAU heating values may be underestimated. The maxima IAU(T) over the ocean of $\sim 0.1 \text{ K d}^{-1}$ are close to the maxima found over the whole tropical or subtropical oceans.

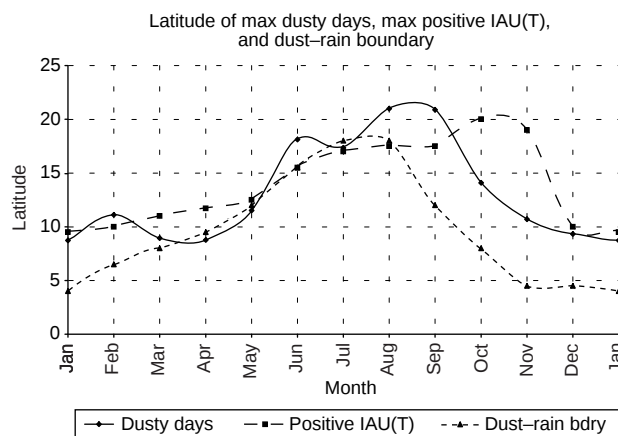


Figure 2 The latitudinal variation of both the maximum IAU(T) (dashed line) and the latitude of maximum dusty days (solid line). The dotted line shows the latitudinal transition of the boundary between the shallow monsoon zone and the zone with convective rainfall (zones B and C1 in ref. 16).

corresponds to about 1.5–3.5 km altitude, typical for Saharan dust layers over the Atlantic¹⁴. In Fig. 1 (left), shading indicates positive values over the ocean, which correspond to a warmer atmosphere compared with the model first-guess prediction.

The monthly variation of both the latitude of maximum positive IAU(*T*) over ocean and the latitude of maximum number of dusty days is shown in Fig. 2. Latitudes of maxima were defined for both curves according to the zonal means over the body of water in the domain 34–4°W. The correlation between the two curves is 0.67 and increases to 0.85 if the latitude of maximum heating is shifted by one month. It suggests that the maximum atmospheric heating in the dust layer lags the maximum in the number of dusty days. The delay may be due to a different process that adds to the IAU(*T*) seasonal cycle, such as thermal heating of clouds. Indeed, monthly correlation of ISCCP (International Satellite Cloud Climatology Project)¹⁵ low cloudiness and the IAU (not shown) is high during the July–November period, which is the same time as when the dust is weaker (Fig. 3).

There is a strong relationship between the latitudinal transition of the African monsoon boundary and that of the dust (Fig. 2). The dotted line shows the latitudinal transition of the boundary between the shallow monsoon zone and the zone with convective rainfall (zones B and C1 in ref. 16). A similarly high correlation was also found with the latitudinal transition of the Harmattan monsoon sharp wind change from the easterly to westerly sector, but this line (not shown), as expected, precedes both the dust and the IAU; the monsoon rain follows the maximum dust line from the south (Fig. 2) because the rain washes out most of the dust.

There is also a high correlation, *r*, between the heating rate (K per day) and the number of dust days, *r* = 0.58, as shown in Fig. 3. This relation is given by

$$\text{IAU}(T) = 0.027 + 7.1 \times 10^{-3} \times N_{\text{dust}} \quad (1)$$

where *N*_{dust} is the average number of dusty days per box, calculated as the total number of dusty days over the domain divided by the number of grid points. Hence, there is an unexplained heating, that is, heating not due to dust, of 0.027 K per day plus a heating due to dust of up to 0.035 K per day for March. The latter value fits well with the independent most recent model estimations for the longitudinal average latitude belt of 10–30°N (ref. 3). The slope constant in the above equation means that an averaged dusty day causes a monthly averaged heating rate of 7.1×10^{-3} K per day, or

0.21 K per month ($30 \text{ d m}^{-1} \times 0.0071 \text{ K d}^{-1}$). However, if we assume that the heating happens only during the dusty-day event, and not during the other days, then this monthly heating rate is really an average of *h*, the heating rate during the single dusty day, and 29 zero-value heating rates during the other days. Therefore, *h*, the heating rate during an averaged dusty day, is 0.21 K per day. Maximum heating values over land of 0.5 K per day within the source of dust over the west Sahara also fit very well the model simulations for Saharan dust maximum heating values¹⁷. The averaged total number of dusty days in a year is 28.6 d yr^{-1} , corresponding to an average heating of 6.0 K yr^{-1} for the full year ($28.6 \text{ d yr}^{-1} \times 0.21 \text{ K d}^{-1}$).

The atmospheric response to the dust forcing was also calculated in an alternative and more powerful approach by comparing each dust pixel to the corresponding IAU pixel. Figure 4 is the resulting jittered scatter plot for 2,496 points: 208 points for each month. In contrast to Fig. 3, this comparison contains both the dust/IAU patterns' resemblance for each month and the intermonthly variation. The linear correlation coefficient is 0.41, increasing to 0.48 when the no-dust pixels are omitted. It is interesting to note that the dust forcing deduced from the slope of the curve based on the moving locally weighted regression is 0.19 K d^{-1} (slope for the 3–14 d period multiplied by 30 d), which is quite close to the aforementioned value of 0.21 K d^{-1} derived from the monthly averages. Also, the unexplained bias is reduced from 0.027 K d^{-1} to 0.0062 K d^{-1} , as calculated for a linear regression. The moving locally weighted regression, the kernel regression¹⁸, shows near saturation for a large number of dusty days, that is, when it exceeds 14–15 d. Such a saturation may be expected because the increased number of dusty days also probably reflects an increase in the average dust concentrations.

A similar scatter plot for the relation between IAU and low-level cloudiness from the ISCCP data sets¹⁶ does not reveal any significant correlation, the value being only 0.11. This, however, does not reflect the high positive cloud/IAU correlations we have found for the intermonthly variations because of the high anticorrelated cloud/IAU patterns for December–June (not shown). These complex IAU/cloud relationships are strongly dependent on the cloud parameterizations and conclusive results are therefore impossible to obtain. This is in contrast to the IAU/dust relationship as no dust parameterization was incorporated in the assimilation system.

An independent assessment of the relationship between the

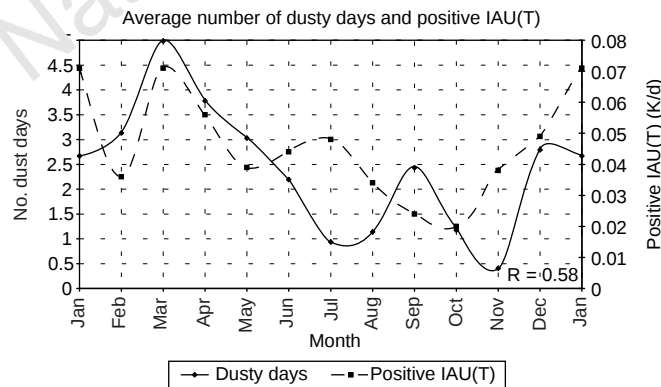


Figure 3 The monthly heating rate in degrees Kelvin per day, as deduced from IAU(*T*) and the average number of dusty days. The regression relation is given by equation (1), where *N*_{dust} is the number of dusty days with correlation between the two curves of *r* = 0.58. The unexplained heating of 0.027 K d^{-1} may be due to other deficiencies such as the model's inability to simulate fully the warmth of the Saharan Air Layer. Most of the SAL heating effect on the IAU, however, is below the dust layer.

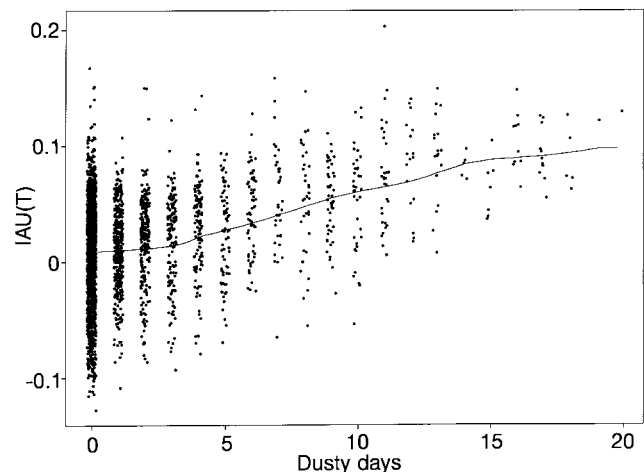


Figure 4 Jittered scatter plot for IAU(*T*) versus number of dusty days for each pixel; 208 points for each month total of 2,496 points. A moving locally weighted regression line is plotted and shows the saturation at high number of dusty days. 'Jittered' describes a random spreading of the many points sharing the same value in the close vicinity of the real value, so that the visual information in the graph correctly reflects the number of data points.

aforementioned heating rate and dust radiative forcing can be gained from comparison of the rate of absorption of solar radiation by the dust layer. For example, our measurements from Cape Verde (16.7° N, 23° W) for 1994–95 suggest an average optical thickness of 0.35 for the dust layer. The heating rate in the 650–850 hPa layer for $\tau = 0.35$ can be estimated from ref. 19 (see their Figs 12, 13) by interpolating values given for $\tau = 0.2$ and $\tau = 0.5$. This procedure gives heating rates of 0.1–0.2 K d⁻¹ for the cloud-free ocean case, and up to 0.3–0.5 K d⁻¹ in the cloud case, which fits well with our estimations for an average dust event of 0.21 K d⁻¹ or 0.19 K d⁻¹ (derived from Figs 3 and 4, respectively).

Another comparison can be made with most recent modelling results from Tegen *et al.*³, who indicate that the average mean global optical thickness coming from disturbed soil is 0.017. To retrieve our mean value (0.35), we have to assume that dust with $\tau = 0.35$ is covering 5% of the total Earth's surface. As dust is located between 10° and 30° N, approximately one-third of the 10–30° latitude belt would be covered by a dust cloud of $\tau = 0.35$. They further indicate that the longitudinal mean heating between 10° and 30° N is 0.04 K d⁻¹: this would result in a heating rate that is three times larger where dust is located, or 0.12 K d⁻¹.

There is also a dynamical heating effect on the IAU due to the Saharan air layer (SAL). This, however, has a different spatial and temporal signature from dust and concentrates in a lower-altitude layer²⁰. The sporadic heating due to the dust at a particular time and location is determined by the transient local properties of a travelling dust plume—the local dust profile and dust content—whereas that of the SAL is on a larger (synoptic) scale. Therefore, dust plume heating is superposed locally and perhaps semirandomly on that of the SAL. The model is sensitive to the general presence and intensity of the SAL but does not contain dust. The product of the data assimilation system—the IAU—does include the effects of dust and thus can show the much more rapidly variations in local conditions due to the passing of a dust plume. Also, the vertical profiles of both the IAU and the theoretical diabatic heating due to Saharan dust¹⁹ show a sharp increase in the 750–580 hPa layer which is located with the average upper part of the dust layer; the climatological SAL profiles, however, indicate a peak in the heating rate below 900 hPa.

We have also compared the global seasonal IAU(T) for 1987 with equivalent atmospheric optical thicknesses (EAOT) maps from Husar *et al.*⁸: this indicates that the pattern similarities between IAU(T) and EAOT exist not only over the Atlantic Ocean but also over other dusty regions, including the Indian Ocean, the Bay of Bengal and the south China Sea during spring and summer (not shown).

This work indirectly shows the response of the temperature field to the radiative forcing of the dust layer. The resulting heating rate of 6 Kyr⁻¹ within the dust layer at 1.5–3.5 km is in agreement with calculated heating rates^{17,20–22} and therefore may be used for improving atmospheric predictions once dust loading is monitored daily by satellite, as is planned for the MODIS (Moderate Resolution Imaging Spectroradiometer) instrument on the Earth-observing system^{23,24}. □

Received 16 April 1997; accepted 3 July 1998.

1. Charlson, R. *et al.* Climate forcing by anthropogenic aerosols. *Science* **255**, 423–430 (1992).
2. Penner, J. E. *et al.* Quantifying and minimizing uncertainty of climate forcing by anthropogenic aerosols. *Bull. Am. Meteorol. Soc.* **75**, 375–400 (1994).
3. Tegen, I., Laci, A. A. & Fung, I. The influence of mineral aerosols from disturbed soils on the global radiation budget. *Nature* **380**, 419–422 (1996).
4. Arking, A. The radiative effects of clouds and their impact on climate. *Bull. Am. Meteorol. Soc.* **72**, 795–813 (1991).
5. Lindzen, R. S. Some coolness concerning global warming. *Bull. Am. Meteorol. Soc.* **71**, 288–299 (1990).
6. Moulin, C., Lambert, C. E., Dulac, F. & Dayan, U. Control of atmospheric export of dust from North Africa by the North Atlantic Oscillation. *Nature* **387**, 691–694 (1997).
7. Prospero, J. M. *et al.* Temporal variability of summertime ozone and aerosols in the free troposphere over the eastern North Atlantic. *Geophys. Res. Lett.* **22**, 2925–2928 (1995).
8. Husar, R. B., Prospero, J. M. & Stowe, L. L. Characterization of tropospheric aerosols over the oceans with the NOAA advanced very high resolution radiometer optical thickness operational product. *J. Geophys. Res.* **102**, 16889–16910 (1997).
9. Daley, R. *Atmospheric Data Analysis* (Cambridge Univ. Press, New York, 1991).

10. Bloom, S. S., Takacs, L. L., da Silva, A. M. & Ledvina, D. Data assimilation using incremental analysis updates. *Mon. Weath. Rev.* **124**, 1256–1271 (1996).
11. Schubert, S. *et al.* An assimilated data set for Earth Science applications. *Bull. Am. Meteorol. Soc.* **74**, 2331–2342 (1993).
12. Jankovik, I. & Tanre, D. Satellite climatology of Saharan dust outbreaks: Method and preliminary results. *J. Clim.* **15**, 646–656 (1992).
13. Prospero, J. M. & Nees, R. T. Impact of the North African drought and El Niño on the mineral dust in the Barbados trade winds. *Nature* **320**, 735–738 (1986).
14. Westphal, D. L., Toon, O. B. & Carlson, T. N. A two-dimensional numerical investigation of the dynamics and micro-physics of Saharan dust storms. *J. Geophys. Res.* **92**, 3027–3049 (1987).
15. Rossow, R. B. & Schiffer, R. A. ISCCP cloud data products. *Bull. Am. Meteorol. Soc.* **72**, 2–20 (1991).
16. Asnani, G. C. *Tropical Meteorology* (Indian Inst. of Tropical Meteorology, Pune, 1993).
17. Pick, C. *Transport of Desert Aerosols and their Influence on Local Temperature and Motion*. Thesis, Tel Aviv Univ. (1991).
18. Eubank, R. L. *Spine Smoothing and Non-parametric Regression* (Dekker, New York, 1988).
19. Carlson, T. N. & Benjamin, S. G. Radiative heating rates for Saharan dust. *J. Atmos. Sci.* **37**, 193–213 (1980).
20. Karyampudi, V. M. *A Numerical Study of the Evolution, Structure and Energetics of the Saharan Air Layer*. Thesis (Pennsylvania State Univ., 1986).
21. Joseph, J. H. in *Aerosols and their Climatic Effects* (eds Gerber, H. E. & Deepak, A.) 215–226 (Deepak, Hampton, VA, 1984).
22. Alpert, P. & Ganor, E. A jet-stream associated heavy dust storm in the western Mediterranean. *J. Geophys. Res.* **98**, 7339–7349 (1993).
23. Tanré, D., Kaufman, Y. J., Herman, M. & Mattoo, S. Remote sensing of aerosol over oceans using the EOS-MODIS spectral radiances. *J. Geophys. Res.* **102**, 16971–16988 (1997).
24. Kaufman, Y. J. *et al.* Remote sensing of tropospheric aerosol over land from EOS moderate resolution imaging spectroradiometer. *J. Geophys. Res.* **102**, 17051–17068 (1997).
25. Reale, A. L., Chalfant, M. W., Wagoner, R. V., Gardner, T. J. & Casey, L. W. TOVS Operational Sounding Upgrades 1990–1992. *NOAA Tech. Rep. NESDIS 76* (US Dept of Commerce, NOAA NESDIS, Washington DC, 1994).

Acknowledgements. This study was supported by the US–Israel Binational Science Foundation. During some of this work, P.A. held a National Research Council–NASA/GSFC research associateship. Our thanks to Y. Benjamini for statistical help with Fig. 4.

Correspondence should be addressed to P.A. (e-mail: pinhas@cyclone.tau.ac.il).

Evidence for long-term diffuse deformation of the lithosphere of the equatorial Indian Ocean

Richard G. Gordon*, Charles DeMets† & Jean-Yves Royer‡

* Department of Geology & Geophysics, Rice University, Houston, Texas 77005, USA

† Department of Geology & Geophysics, University of Wisconsin, Madison, Wisconsin 53706, USA

‡ Géosciences Azur, CNRS UMR 6526, BP48, 06235 Villefranche-sur-Mer cedex, France

The presence of large earthquakes, east–west-striking folds and thrust faults in sediments, and east–west-striking undulations of wavelength 200 km in topography and gravity shows that the equatorial Indian Ocean is the locus of unusual deformation^{1–8}. This deformation has been interpreted as a diffuse boundary between two tectonic plates^{9–13}. Seismic stratigraphy and deep-sea drilling at two locations in the Bengal fan indicate that the deformation began 7.5–8.0 Myr ago^{3,14,15}. Here, however, we show, using plate reconstructions, that motion across this diffuse oceanic plate boundary began more than 10 Myr earlier than previously inferred and that the amount of north–south convergence across the boundary through the central Indian basin has been significantly greater than the convergence estimated from seismic profiles. The relative plate velocity accommodated across the central Indian basin has varied with time and has been as fast as ~6 mm yr⁻¹—about half the separation rate of Earth's slowest-spreading mid-ocean ridge. The earliest interval of measurable motion, which began more than 18 Myr ago, may coincide with rapid denudation of the Tibetan plateau from ~21 Myr to 15–17 Myr (ref. 16). The present motion across the central Indian basin began no earlier than 11 Myr—following an earlier interval of slower motion from 18 to 11 Myr—and may have begun at ~8 Myr, when the Tibetan plateau is thought to have attained its maximum elevation^{16,17}.

The traditionally defined Indo-Australian plate is a composite plate comprising three distinct plates and many deforming zones that can be usefully treated as diffuse plate boundaries¹³ (Fig. 1). Most studies of distributed lithospheric deformation in the Indian Ocean have focused on the equatorial region south of India. Deformation in this zone accommodates motion mainly between the Indian and Capricorn plates but also between these two plates and the Australian plate¹³ (Fig. 1). Two major unconformities, revealed by an extensive network of seismic profiles, can be traced across the Bengal fan³. The age of the younger unconformity coincides with the onset of the folding and faulting that continues today³. This unconformity, which has been drilled in two places separated by ~1,100 km (Fig. 2), is 7.5–8.0 Myr old^{3,14,15} and may coincide with the onset of normal faulting in the Tibetan plateau at ~8 Myr (ref. 16), when the plateau is believed to have first attained an elevation that equals or exceeds its present elevation^{16,17}. The attainment of maximum elevation may in turn be related to significant climate change at 6–8 Myr, including the intensification of the monsoon over the Arabian Sea^{18,19} and warming, increased aridity, and a change from forests to grasslands in northern Pakistan²⁰. Molnar *et al.*¹⁷ argue that the maximum elevation of the Tibetan plateau was attained only after its lithospheric mantle was convectively destabilized and removed. They further hypothesize that the resulting uplift and increase of potential energy of the plateau provided the large force required to cause folding of the Indo-Australian lithosphere, which occurs only if the force per unit length exceeds $\sim 4 \times 10^{12} \text{ N m}^{-1}$ (ref. 17).

Here we present estimates of the motion of the Indian plate relative to the Capricorn plate since 18 Myr and since 20 Myr, and compare them with each other and with published estimates of the

motion since 3 Myr (ref. 12) and 11 Myr (ref. 21). The reconstructions are determined from 81 crossings of the young end of anomaly 5E (18.3 Myr; ref. 22), 102 crossings of the old end of anomaly 6 (anomaly 6no, 20.1 Myr; ref. 22), 125 crossings of fracture zones near anomaly 5E, and 101 crossings of fracture zones near anomaly 6no flanking the central Indian and Carlsberg ridges. We used the 2-minute gravity grid of Sandwell and Smith²³ to select fracture zone crossings at 15-km intervals along the gravity troughs that we assume mark the axis of the fracture zones. Finite rotations between plates and the uncertainties in the rotations are determined using methods described by Royer and Chang¹¹.

Our plate reconstructions show that there is a large and obvious difference between the rotation of the Indian plate relative to the Capricorn plate since 20 Myr and the rotation since 11 Myr (Fig. 3). We rigorously tested whether India-Capricorn motion since chron 6 (20 Myr) exceeds that since chron 5 (11 Myr) by differencing the

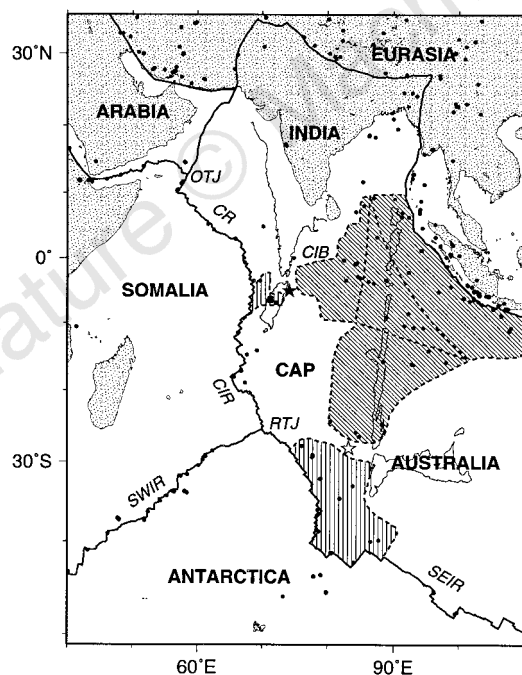


Figure 1 Location map. Small filled circles show the locations of earthquakes with magnitudes ≥ 5.5 . The filled star shows the approximate pole of present rotation between the Indian and Capricorn plates; the open star shows the approximate pole of present rotation between the Capricorn and Australian plates. Vertically striped regions are diffuse plate boundaries across which divergence is accommodated; the diagonally striped region shows diffuse plate boundaries across which convergence is accommodated. Abbreviations: CAP, Capricorn plate; OTJ, Owen triple junction; CR, Carlsberg ridge; CIB, central Indian basin; CIR, central Indian ridge; RTJ, Rodrigues triple junction; SEIR, southeast Indian ridge; SWIR, southwest Indian ridge.

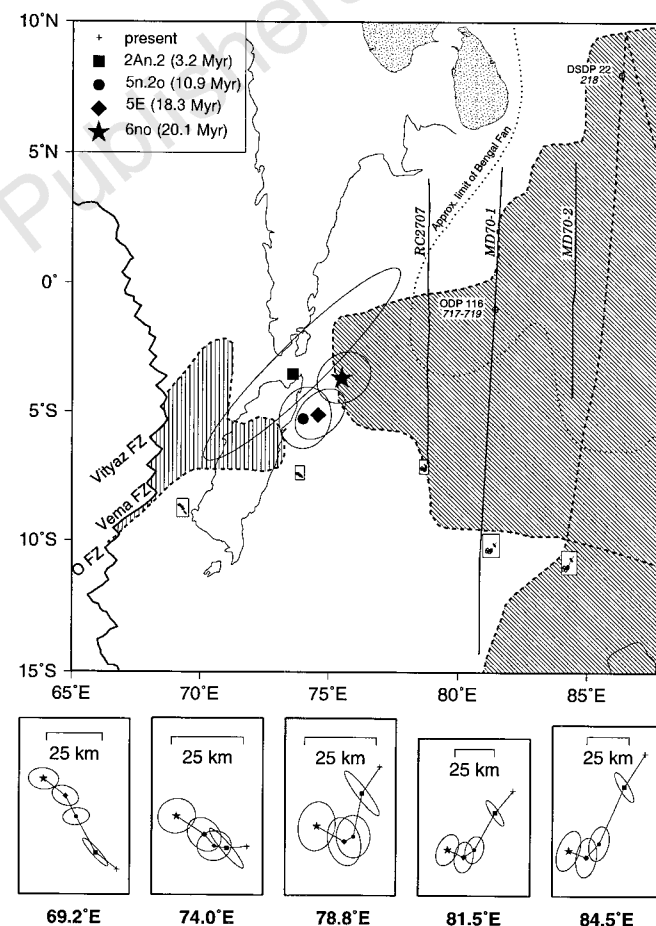


Figure 2 Map of equatorial Indian Ocean showing poles of finite rotation (and 95% confidence regions) between the Indian and Capricorn plates. The three north-south-striking lines labelled RC2707, MD70-1 and MD70-2 show the locations of seismic profiles along which north-south convergence has been estimated⁶⁻⁷. The location of ocean drilling sites (Site 218 from DSDP Leg 22 and Sites 717-719 from ODP Leg 116) are shown by circles with plus signs. Each of the five small rectangles in the top panel shows the reconstructed position of a point on the Capricorn plate relative to an arbitrarily fixed Indian plate; enlarged versions of these rectangles with their reconstructed points and 95% confidence regions are shown below the top panel. The reconstructions for anomaly 2A (3.2 Myr; ref. 22) and the old end of anomaly 5 (anomaly 5n.2o; 10.9 Myr; ref. 22) are modified from DeMets *et al.*¹² and Royer *et al.*²¹, respectively. For each reconstruction, the chron identification, pole latitude ($^{\circ}$ N), pole longitude ($^{\circ}$ E), rotation angle (degrees), and the xx , xy , xz , yy , yz , and zz components of the covariance matrix are given in Table 1.

Table 1 Finite rotations and covariances

Chron	Pole location		Ω (deg.)	Covariance matrix					
	Lat. (° N)	Long. (° E)		xx	xy	xz	yy	yz	zz
India-Somalia									
2An.2	26.23	28.14	-1.226	126.6	221.2	-96.3	488.5	-101.7	124.2
5n.2o	24.40	30.36	-4.349	354.7	744.5	19.2	1675.2	59.3	27.0
5E	24.22	32.25	-7.880	355.0	821.2	0.7	2074.6	29.6	20.9
6no	25.30	30.65	-8.488	551.2	1243.1	30.6	3012.3	83.0	35.8
Capricorn-Somalia									
2An.2	14.57	48.74	-1.927	75.4	148.7	-61.4	545.7	-83.0	63.1
5n.2o	13.67	47.07	-6.319	327.6	794.7	-235.9	2149.9	-615.4	199.6
5E	15.92	44.44	-10.239	557.6	1317.4	-340.6	3487.2	-820.5	241.8
6no	17.44	43.26	-10.861	395.6	968.9	-259.4	2774.2	-642.1	192.1
India-Capricorn									
2An.2	-3.54	73.60	0.922	389.7	436.5	-272.3	819.3	-263.3	214.4
5n.2o	-5.27	73.99	2.625	696.3	1555.4	-218.3	3816.4	-546.4	221.2
5E	-5.12	74.59	3.230	936.0	2162.6	-337.6	5550.9	-769.1	250.1
6no	-3.69	75.50	3.345	968.6	2235.1	-228.3	5775.4	-542.4	217.2

Ω is rotation angle; positive rotation angles are right-handed. Each rotation reconstructs the first-named plate relative to the second. The last six columns give the indicated elements of the covariance matrix in units of 10^{-8} rad^2 with the x, y and z axes being parallel to (0°N, 0°E), (0°N, 90°E) and 90°N, respectively.

India-Capricorn rotations for these two ages while propagating the errors¹¹. A χ^2 test comparing this rotation with the null rotation gives a value of χ^2 of 78.3 with 3 degrees of freedom. The probability of obtaining a value this large or larger by chance if there were no motion between the Indian and Capricorn plates before 11 Myr is 1 in 7×10^{16} . An even greater minimum age of onset of motion between the Indian and Capricorn plates is obtained by differencing India-Capricorn rotations since the young end of chron 5E (18 Myr) and since the old end of chron 6 (20 Myr). A χ^2 test comparing this difference in rotations with the null rotation gives a value of χ^2 of 13.2 with 3 degrees of freedom. The probability of obtaining a value this large or larger by chance if there were no motion between the Indian and Capricorn plates before 18 Myr is 1 in 4×10^3 . Thus motion began before 18 Myr, at least 10 Myr before the age of onset of deformation (7.5–8.0 Myr) inferred previously.

Our minimum age for the onset of India-Capricorn motion implies that deformation began somewhere in the equatorial Indian Ocean other than where the Bengal fan was drilled, or that the style of deformation changed at 7.5–8.0 Myr, or both. Our results are consistent with an onset of deformation of Indo-Australian lithosphere at ~ 20 Myr, coincident with the onset of rapid denudation of the Tibetan plateau from ~ 21 to ~ 15 –17 Myr as indicated by thermochronometry^{24,25}, Sr isotopic evolution of sea water²⁶, and the sedimentary record of the Siwalik Group and the Bengal fan^{27–29}. Our reconstructions indicate that motion across the central Indian basin was faster ($\sim 6 \text{ mm yr}^{-1}$ to the ESE for a point now at 10°S, 81.5°E on the Capricorn plate) from 20 to 18 Myr, possibly coinciding with the episode of rapid denudation of the Tibetan plateau, than from 18 to 11 Myr when motion was slow ($\sim 1 \text{ mm yr}^{-1}$ to the NE; Fig. 2). Our results are, however, also consistent with Indo-Australian deformation having started even earlier.

The Indo-Australian plate is believed to have come into existence shortly after chron 20 (43 Myr) when spreading between the Indian and Australian plates ceased in the Wharton basin³⁰. Therefore, the Indo-Australian plate existed as a single nearly rigid plate for at most ~ 20 Myr and may never have existed at all. These alternatives may be testable by accurately reconstructing the motion between the Indian and Capricorn plates at one or more ages long before 20 Myr. In any event, our results show that the deformation of equatorial Indian Ocean lithosphere is not ephemeral, but long-lived.

The north–south component of convergence across the central Indian basin has been estimated from the throws and dips of thrust faults observed on north–south seismic profiles^{5–7} (Figs 2, 4). The plate reconstructions indicate that the convergence since both chron

5E and chron 6 exceeds that estimated from each seismic profile (Fig. 4). If the seismic profiles record all the north–south shortening along them, as has been assumed in previous work, then our results would be inconsistent with the convergence estimates from the profiles. If instead we assume that the estimates from seismic profiles are only measuring the convergence since 7.5–8.0 Myr, then these estimates may be consistent with those from the plate reconstructions. Ideally the estimates from seismic profiles would be compared with those for plate motion since 7.5–8.0 Myr. Unfortunately there is no reconstruction available for this age. Instead we compare the convergence estimated from each seismic profile to those since 11 Myr. The ratio of the convergence estimated from seismic profiles to the convergence estimated from chron 5 plate reconstructions is 0.45, 0.77 and 0.82 along 78.8°E, 81.5°E and 84.5°E, respectively. The last two ratios are similar to the ratio, 0.69–0.73, of the assumed age of onset of the deformation estimated from the seismic profiles (7.5–8.0 Myr) to the old end of chron 5 (10.9 Myr). This leaves the discrepancy between the convergence estimates along 78.8°E to be explained, perhaps because the faults are listric and perhaps because of motion out of the vertical plane of the seismic profile.

Is some tectonic event other than the onset of motion between the Capricorn and Indian plates recorded by the seismic stratigraphy and ocean drilling? Perhaps it is an acceleration of motion across the central Indian basin and the onset of a particular mode of deformation, the lithosphere-scale folding and associated thrust faulting and small-scale folding that presumably continue today. The rate of motion accommodated across the central Indian basin was slow ($\sim 1 \text{ mm yr}^{-1}$) from 18 to 11 Myr, but increased to $\sim 4 \text{ mm yr}^{-1}$ to the NNE (near 10°S, 81.5°E) between 11 Myr and 3 Myr, and further increased to $\sim 6 \text{ mm yr}^{-1}$ to the NNE since 3 Myr. Thus, the constraints from the plate reconstructions are consistent with plate motion across the central Indian basin having accelerated at ~ 8 Myr, when the Tibetan plateau is believed to have attained its maximum elevation. Additional plate reconstructions for ages between 11 Myr and 3 Myr could bracket the timing more closely. In any event, the force that deformed the diffuse plate boundary before ~ 8 Myr must have been too small, in the wrong direction, or both, to cause lithosphere-scale folding. Our reconstructions indicate that, on average, points on the Capricorn plate in the southern central Indian basin moved mainly to the east from ~ 20 to ~ 11 Myr, and moved to the NNE since that time (see the points at 78.8°E, 81.5°E and 84.5°E in Fig. 2). If the boundary before 11 Myr was also an east–west-striking zone, as is the present boundary, the deformation that occurred before 11 Myr ago was east–west simple shearing. □

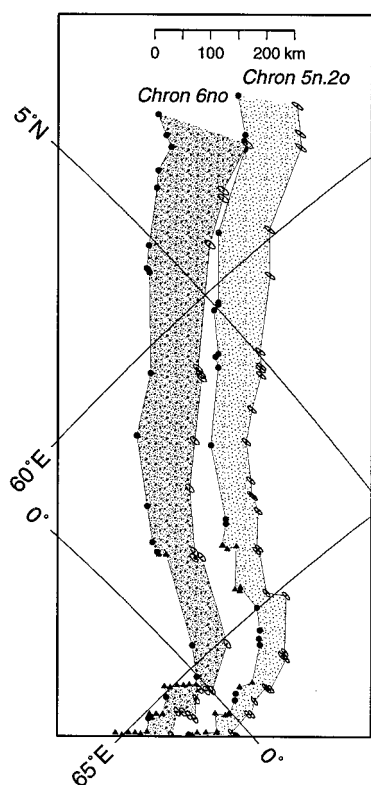


Figure 3 Reconstructions of magnetic anomalies 6 and 5 along the Carlsberg ridge. Locations of crossings of the old end of magnetic anomalies 6 (anomaly 6no; left) and 5 (anomaly 5n.2o; right) on the Somalian plate, which is arbitrarily held fixed, are shown as filled circles; associated fracture crossings are shown as filled triangles. A thin line connects the Somalian 6no crossings and another thin line connects the Somalian 5n.2o crossings. Coeval Indian plate magnetic anomaly and fracture zone crossings, which are shown by their 95% confidence ellipses, have been reconstructed using the appropriate rotation that best fits the Capricorn plate to the Somalian plate for either chron 6no or chron 5n.2o. Each set of Indian crossings is also connected by a thin solid line. The stippled region on the left shows the gap between the Somalian plate and Indian plate crossings for chron 6no, whereas the stippled region on the right shows the gap for chron 5n.2o. If India and Capricorn were part of a single rigid plate, the Somalian and Indian crossings would coincide in location within their uncertainties. The gaps shown here are caused by the neglect of the motion between the Indian and Capricorn plates. If the motion between the Indian and Capricorn plates occurred entirely since chron 5n.2o (10.9 Myr; ref. 22), as would be the case if motion began 7.5–8.0 Myr as inferred from seismic stratigraphy and drilling in the Bengal fan¹⁵, the chron 6no (20.1 Myr; ref. 22) gap would be the same within uncertainties as the chron 5n.2o gap. The chron 6no gap is significantly larger, however, than the chron 5n.2o gap. The former is as large as ~140 km along the northwestern Carlsberg ridge (top of panel), whereas the latter is only as large as ~100 km. The figure is an oblique Mercator projection about the India-Somalia stage pole rotation from chron 6no to chron 5n.2o.

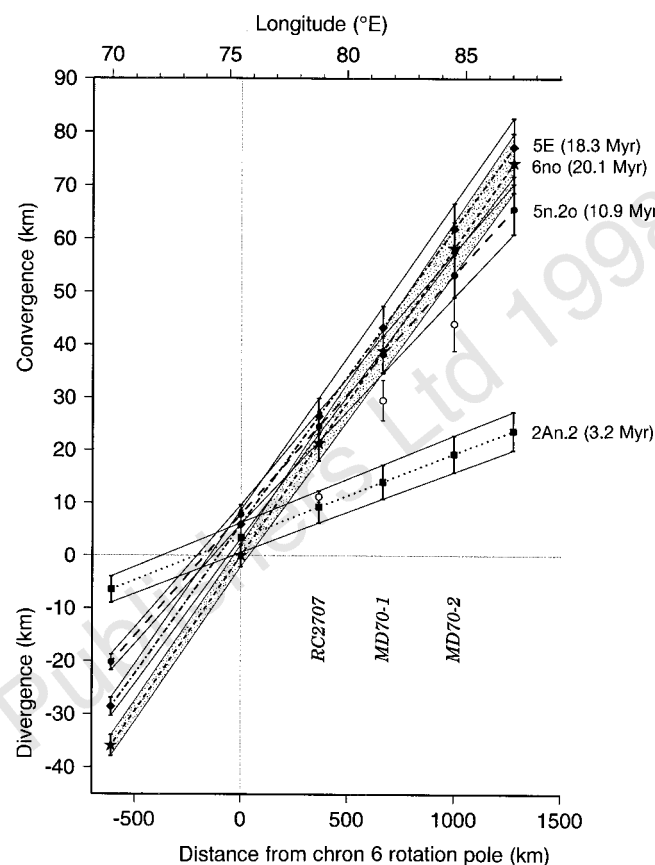


Figure 4 Estimated north-south convergence or divergence between the Indian and Capricorn plates. Convergence or divergence inferred from plate reconstructions (filled symbols) are shown for lines that are now north-south in the equatorial Indian Ocean since the middle of chron 2A¹² (3.2 Myr), the old end of chron 5²¹ (chron 5n.2o; 10.9 Myr), the young end of chron 5E (18.3 Myr), and the old end of chron 6 (chron 6no; 20.1 Myr). Also shown are estimates of convergence from north-south seismic lines along 78.8°E, 81.5°E, and 84.5°E (open circles)⁵⁻⁷. Vertical bars show $\pm 1\sigma$ errors. The convergence since 3.2 Myr is less than estimated along each seismic profile, but insignificantly so along 78.8°E, whereas the convergence since 10.9 Myr is greater than that along each profile, significantly so only along 78.8°E²¹. The convergence since chron 5E and since chron 6 exceeds that estimated along every seismic profile: 27.1 ± 6.7 km (5E) and 22.5 ± 6.5 km (6) versus 11 ± 2 km (ref. 7) along 78.8°E, 43.8 ± 8.0 km (5E) and 39.9 ± 8.0 km (6) versus 29.5 ± 7.5 km⁵ along 81.5°E, and 62.3 ± 9.4 km (5E) and 59.1 ± 9.6 km (6) versus 44 ± 10 km (ref. 6) along 84.5°E. Averaging the north-south convergence estimated from the 5E and 6 plate reconstructions and then subtracting the north-south convergence estimated from seismic profiles gives the following differences: 13.6 ± 5.0 km along 78.8°E, 12.3 ± 9.4 km along 81.5°E, and 16.7 ± 12.0 km along 84.5°E. (Here all values after $\pm$ symbols are 95% confidence limits.)

Received 19 November 1997; accepted 11 June 1998.

- Stein, S. & Okal, E. A. Seismicity and tectonics of the Ninetyeast Ridge area: Evidence for internal deformation of the Indian plate. *J. Geophys. Res.* **83**, 2233–2245 (1978).
- Weissel, J. K., Anderson, R. N. & Geller, C. A. Deformation of the Indo-Australian plate. *Nature* **287**, 284–291 (1980).
- Curry, J. R. & Munasinghe, T. Timing of intraplate deformation, northeastern Indian Ocean. *Earth Planet. Sci. Lett.* **94**, 71–77 (1989).
- Bull, J. M. & Scrutton, R. A. Seismic reflection images of intraplate deformation, Central Indian Ocean, and their tectonic significance. *J. Geol. Soc. Lond.* **149**, 955–966 (1992).
- Chamot-Rooke, N., Jestin, F., de Voogd, B. & Phèdre Working Group Intraplate shortening in the central Indian Ocean determined from a 2100-km-long north-south deep seismic reflection profile. *Geology* **21**, 1043–1046 (1993).
- Jestin, F. *Cinématique rigide et Déformations dans la Jonction Triple Afar et dans le Bassin Indien Central*. Thesis, Univ. Pierre et Marie Curie (Paris 6) (1994).
- Van Orman, J., Cochran, J. R., Weissel, J. K. & Jestin, F. Distribution of shortening between the Indian and Australian plates in the central Indian Ocean. *Earth Planet. Sci. Lett.* **133**, 35–46 (1995).
- McAdoo, D. C. & Sandwell, D. T. Folding of oceanic lithosphere. *J. Geophys. Res.* **90**, 8563–8569 (1985).

- Wiens, D. A. *et al.* A diffuse plate boundary model for Indian Ocean tectonics. *Geophys. Res. Lett.* **12**, 429–432 (1985).
- Gordon, R. G., DeMets, C. & Argus, D. F. Kinematic constraints on distributed lithospheric deformation in the equatorial Indian Ocean from present motion between the Australian and Indian plates. *Tectonics* **9**, 409–422 (1990).
- Royer, J.-Y. & Chang, T. Evidence for relative motions between the Indian and Australian plates during the last 20 Myr from plate tectonic reconstructions: Implications for the deformation of the Indo-Australian plate. *J. Geophys. Res.* **96**, 11779–11802 (1991).
- DeMets, C., Gordon, R. G. & Vogt, P. Location of the Africa-Australia-India triple junction and motion between the Australian and Indian plates: Results from an aeromagnetic investigation of the central Indian and Carlsberg ridges. *Geophys. J. Int.* **119**, 893–930 (1994).
- Royer, J.-Y. & Gordon, R. G. The motion and boundary between the Capricorn and Australian plates. *Science* **277**, 1268–1274 (1997).
- Moore, D. G., Curry, J. R., Raitt, R. W. & Emmel, F. J. Stratigraphic-seismic section correlations and implications to Bengal Fan history. *Initial Rep. DSDP* **22**, 403–412 (1974).
- Cochran, J., Stow, D. & Leg 116 shipboard scientific party Collisions in the Indian Ocean. *Nature* **330**, 519–521 (1987).
- Harrison, T. M., Copeland, P., Kidd, W. S. F. & Yin, A. Raising Tibet. *Science* **255**, 1663–1670 (1992).

17. Molnar, P., England, P. & Martinod, J. Mantle dynamics, uplift of the Tibetan Plateau, and the Indian Monsoon. *Rev. Geophys.* **31**, 357–396 (1993).
18. Kroon, D., Steens, T. & Troelstra, S. R. Onset of monsoonal related upwelling in the western Arabian Sea as revealed by planktonic foraminifers. *Proc. ODP Sci. Res.* **117**, 257–263 (1991).
19. Prell, W. L., Murray, D. W., Clemens, S. C. & Anderson, D. M. in *Synthesis of Results from Scientific Drilling in the Indian Ocean* (ed. Duncan, R. A. et al.) 447–469 (Geophys. Monogr. Vol. 70, Am. Geophys. Union, Washington DC, 1992).
20. Quade, J. T., Cerling, T. E. & Bowman, J. R. Development of Asian monsoon revealed by marked ecological shift during the latest Miocene in northern Pakistan. *Nature* **342**, 163–166 (1989).
21. Royer, J.-Y., Gordon, R. G., DeMets, C. & Vogt, P. R. New limits on the motion between India and Australia since chron 5 (11 Ma) and implications for lithospheric deformation in the equatorial Indian Ocean. *Geophys. J. Int.* **129**, 41–74 (1997).
22. Cande, S. C. & Kent, D. V. Revised calibration of the geomagnetic polarity timescale for the Late Cretaceous and Cenozoic. *J. Geophys. Res.* **100**, 6093–6095 (1995).
23. Sandwell, D. T. & Smith, W. H. F. Marine gravity from Geosat and ERS-1 satellite altimetry. *J. Geophys. Res.* **102**, 10039–10054 (1997).
24. Copeland, P., Harrison, T. M., Kidd, W. S. F., Xu, R. & Zhang, Y. Rapid early Miocene acceleration of uplift in the Gangdese belt, Xixang (southern Tibet) and its bearing on accommodation mechanisms of the India-Asia collision. *Earth Planet. Sci. Lett.* **86**, 240–252 (1987).
25. Richter, F. M., Lovera, O. M., Harrison, T. M. & Copeland, P. Tibetan tectonics from $^{40}\text{Ar}/^{39}\text{Ar}$ analysis of a single K-feldspar sample. *Earth Planet. Sci. Lett.* **105**, 266–278 (1991).
26. Richter, F. M., Rowley, D. B. & DePaolo, D. J. Sr isotope evolution of seawater: the role of tectonics. *Earth Planet. Sci. Lett.* **109**, 11–23 (1992).
27. Honegger, K. et al. Magmatism and metamorphism in the Laddakh Himalayas (the Indus-Tsangpo suture zone). *Earth Planet. Sci. Lett.* **60**, 253–292 (1982).
28. Copeland, P. & Harrison, T. M. Episodic rapid uplift of the Himalayas revealed by $^{40}\text{Ar}/^{39}\text{Ar}$ analysis of detrital K-feldspar and muscovite, Bengal fan. *Geology* **18**, 354–357 (1990).
29. Stow, D. A. V. et al. Sediment facies and processes of the distal Bengal Fan, leg 116. *Proc. ODP Sci. Res.* **116**, 377–396 (1990).
30. Liu, C. S., Curry, J. R. & McDonald, J. M. New constraints on the tectonic evolution of the eastern Indian Ocean. *Earth Planet. Sci. Lett.* **65**, 331–342 (1983).

Acknowledgements. We thank P. Vogt and his colleagues at the US Naval Research Laboratory for their help in collecting many of the critical data used in this study. We also thank T. Henstock, S. Cande and D. Rowley for comments on the manuscript. This work was supported by the US NSF and by CNRS.

Correspondence and requests for materials should be addressed to R.G.G. (e-mail: rgg@rice.edu).

Postcranial pneumatization in *Archaeopteryx*

Brooks B. Britt*, Peter J. Makovicky†, Jacques Gauthier‡ & Niels Bonde§

* Museum of Western Colorado, PO Box 20000, Grand Junction, Colorado 81502, USA

† Department of Vertebrate Paleontology, American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024-5192, USA

‡ Department of Geology and Geophysics, Yale University, PO Box 208109, New Haven, Connecticut 06520-8109, USA

§ Institut for Historisk Geologi og Palæontologi, Københavns Universitet, 1350 København K, Denmark

Pneumatization of the postcranial skeleton by the lungs is thought to be a hallmark of the avian skeleton, and to be an adaptation for flight by reducing weight. Pneumatic features have, however, remained elusive in the primitive avialan *Archaeopteryx lithographica*. The hollow long bones of *Archaeopteryx* were first interpreted to be pneumatized¹, but this interpretation was later rejected because of an absence of pneumatic foramina in these bones that connect their interiors with the respiratory system^{2–6}. Pneumatic features have also been recognized in the axial skeleton of many non-avian theropod dinosaurs (and some other archosaurs of the bird clade). The purported lack of postcranial pneumatic features in *Archaeopteryx* has been interpreted as a primitive condition of avialans; this raises doubts about the homology between postcranial pneumatic features of birds and non-avian theropods⁷. Here we re-examine two specimens of *Archaeopteryx*. These specimens show evidence of vertebral pneumaticity in the cervical and anterior thoracic vertebrae, thus confirming the phylogenetic continuity between the pneumatic systems of non-avian theropods and living birds.

Pneumatic foramina are present at the anterior end of the centrum in the first four exposed neck vertebrae on the main slab of the Berlin specimen of *Archaeopteryx lithographica* (MB 1880/81.4598; Fig. 1). We interpret these vertebrae as corresponding to

cervicals 3–6, although they have also been identified as the axis and cervicals 3–5 (ref. 8). The pneumatic foramina lie posterodorsal to the lower attachment of the cervical rib (parapophysis). The foramen of the fifth cervical is best preserved and displays a complete rim, whereas only the dorsal rim and posterior rim are preserved on the fourth and sixth cervical vertebrae, respectively. These latter two vertebrae are more poorly preserved than is the fifth cervical. On the third cervical, the pneumatic foramen is relatively small and inconspicuous. The preserved edges of the foramina are smooth, rounded, and finished, showing that they are neither preparation nor preservation artefacts. All of the visible foramina are filled with microcrystalline calcite, which has a similar colour to the preserved bone, albeit lighter in hue. This calcite infill, and an emphasis on humeral rather than vertebral pneumatization^{2,4}, may have contributed to the earlier oversight of these foramina. Dimensions of the foramina and the vertebrae on which they occur are given in Table 1.

Poor preservation of the remaining cervical vertebrae prevents determination of the posterior extent of centrum pneumatization in the Berlin specimen, but a circular pneumatic foramen is present on the first vertebra to bear a thoracic rib in the Eichstätt specimen (JM 1; Fig. 2). Most of the rim is well-preserved, although the dorsal edge of the foramen is damaged. As in the Berlin specimen, the foramen interior is partially filled with calcite. We did not observe pneumatic openings in the centra of the posterior trunk vertebrae of *Archaeopteryx*, and we could not corroborate a previous claim⁹ that the trunk vertebrae of the Berlin and London specimens bear pleurocoels, a term sometimes, but not always, applied to pneumatic openings in vertebral centra. These pleurocoels^{6,10} may instead correspond to the 'lateral excavations' on one of the posterior trunk vertebrae of the London specimen⁴. Similar depressions are present in the Berlin *Archaeopteryx*, but they are non-invasive.

A small foramen is visible between the two rib heads of the exposed medial surface of the last cervical rib in the Eichstätt

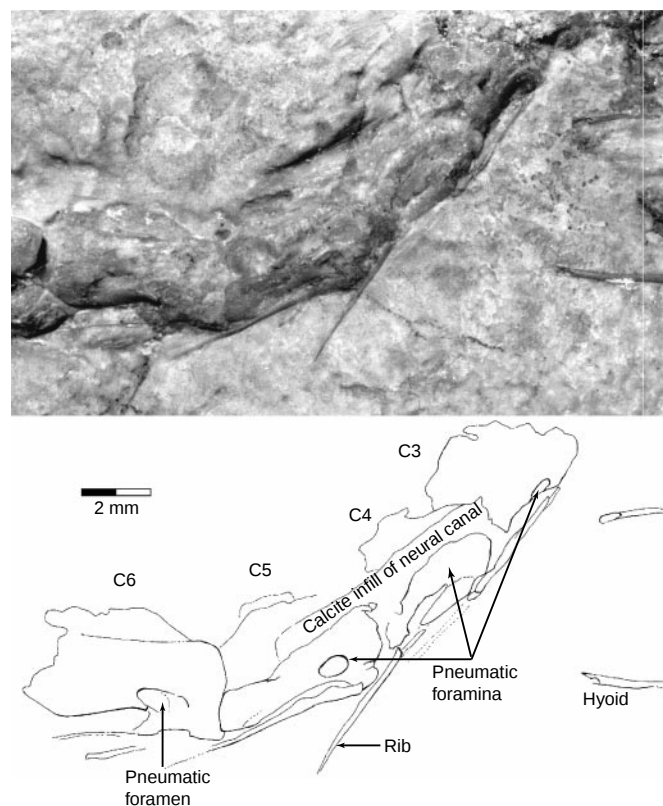


Figure 1 Cervicals 3–6 (C_{3–6}) of the Berlin specimen of *Archaeopteryx lithographica* (HMN 1880. 81.4598) as exposed on the main slab.

Table 1 Measurements of maximum pneumatic foramen dimension and centrum length in *Archaeopteryx*, *Baptornis advenus* and *Alligator*

Taxon and specimen/ vertebral position	Centrum length in mm (ref. 24)	Maximum foramen diameter (mm)	Ratio (centrum: foramen)
Berlin <i>Archaeopteryx</i> /C3	5.0	0.5	10
Berlin <i>Archaeopteryx</i> /C4	6.5	1.1*	5.9
Berlin <i>Archaeopteryx</i> /C5	9.3	1.8	5.2
Berlin <i>Archaeopteryx</i> /C6	9.5	1.4*	6.8
Eichstätt <i>Archaeopteryx</i> /T1	3.5	1.0	3.5
<i>Baptornis</i> (AMNH 5101)/cervical	20.00**	0.33	60.6
<i>Baptornis</i> (AMNH 5101)/cervical	24.40	0.47	51.9
<i>Baptornis</i> (AMNH 5101)/cervical	30.80	0.72	42.8
<i>Baptornis</i> (AMNH 5101)/cervical	20.70**	1.05	19.2
<i>Baptornis</i> (AMNH 5101)/anterior thoracic	22.35	0.47	47.6
<i>Baptornis</i> (AMNH 5101)/anterior thoracic	19.74	0.91	21.7
Juvenile <i>Alligator</i> /axis	5.85	0.34	17.2
Juvenile <i>Alligator</i> /C3	5.92	0.41	14.1
Juvenile <i>Alligator</i> /C4	5.86	0.25	23.4
Juvenile <i>Alligator</i> /C5	5.65	0.41	13.8
Juvenile <i>Alligator</i> /C6	5.46	0.65	8.4
Juvenile <i>Alligator</i> /T1	4.94	0.30	16.5
Juvenile <i>Alligator</i> /T2	5.29	0.42	12.6
Juvenile <i>Alligator</i> /T3	5.57	0.42	13.3
Juvenile <i>Alligator</i> /T4	5.83	0.40	14.6

Comparative measurements of maximum pneumatic foramen diameter and centrum length in *Archaeopteryx*, and maximum nutrient foramen and centrum length in a specimen of *Baptornis advenus* (AMNH 5101) and an embryo of *Alligator*. Only nutrient foramina on the sides of the centra were measured in the latter two specimens. For vertebrae with more than one nutrient foramen per side, the maximum diameter of the largest foramen (left or right) is given. Measurements for *Archaeopteryx* were taken using a millimetre scale ocular on a binocular microscope, or from high magnifications of our photographs. The *Baptornis* and *Alligator* specimens were measured using a digital stage with a microscope. C, cervical vertebra; T, thoracic vertebra. *, Breakage that affects maximum measurement; **, significant amount of breakage.

specimen (Fig. 2). Furthermore, fractured rib shafts in both the Berlin and Eichstätt specimens show that the rib interiors are hollow. The thoracic ribs of living birds are hollow and the internal space is occupied either by diverticula of the cervical air sacs, or by diverticula of the parabronchial lungs, that enter the rib medially between the two heads¹⁰. Such openings are known from the ribs of some non-avian theropods¹¹.

Nutrient foramina are present on the lateral surfaces of the centrum in archosaurs ancestrally, but they are minuscule in size. In birds, cervical air-sac diverticula invade the neck vertebrae late in development^{11,12}, either close to or through these nutrient foramina^{11,13}, forming substantial fenestrae in the process. The pneumatic foramina of advanced theropods, including birds, are proportionally much larger than the primitive nutrient foramina

seen in non-pneumatic vertebrae (Table 1), including apneumatic bird vertebrae (Fig. 3), and also in juvenile crocodylians in which the nutrient foramina are relatively larger than in adults.

The foramina visible in the Berlin and Eichstätt specimens of *Archaeopteryx* are topographically and structurally consistent with the pneumatic openings on the centra of other tetanuran theropods, including extant birds (Aves). In most tetanuran theropods, such as *Compsognathus*, *Allosaurus* and *Ornithomimus*, pneumatic foramina are present on the centra from the axis to the most anterior thoracic vertebrae, a pattern apparently also present in *Archaeopteryx*. Some theropods, including dromaeosaurs, tyrannosaurs and the sister group of *Archaeopteryx*, *Rahonavis*¹⁴, bear pneumatic openings on all presacral vertebrae (except for the atlas), and oviraptors have them on all vertebrae except for those

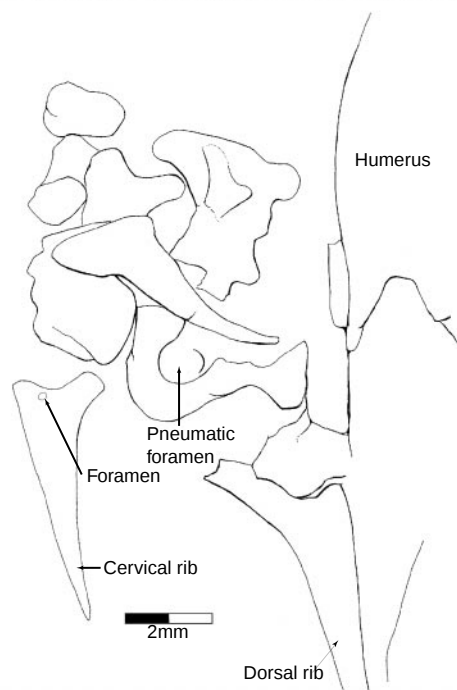
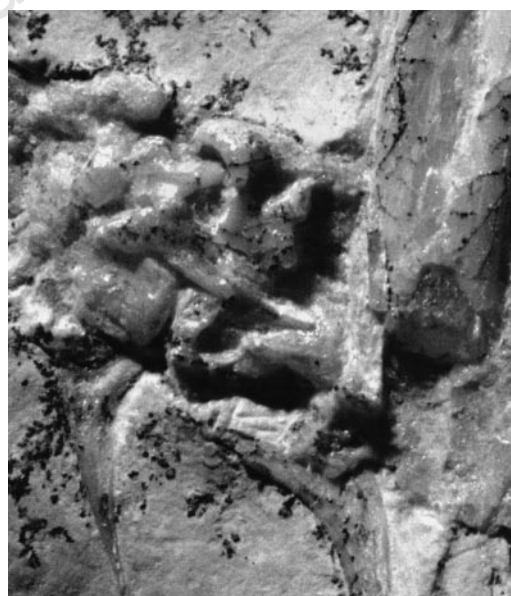


Figure 2 Cervicothoracic transitional region of the Eichstätt specimen of *Archaeopteryx lithographica* (JM 1) as exposed on the underlying slab.

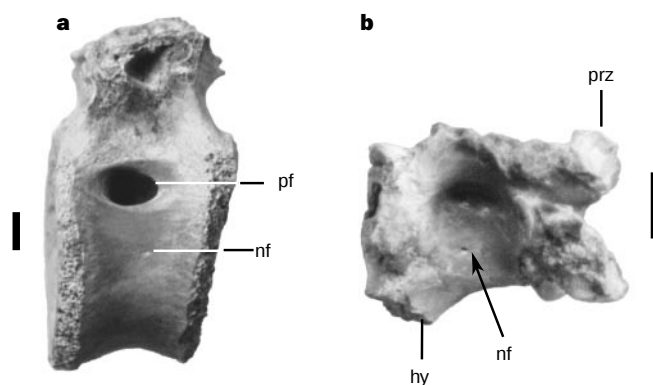


Figure 3 Thoracic vertebrae of **a**, dromaeosaurid theropod (AMNH 21893) and **b**, *Baptornis advenus* (AMNH 5101) in lateral view, showing the relative size difference between a pneumatic foramen (pf) as seen in the dromaeosaur and a non-pneumatic nutrient foramen (nf) present in *Baptornis*. *Baptornis* is a Cretaceous diving bird with apneumatic vertebrae. hy, hypapophysis; prz, prezygapophysis. Scale bars, 5 mm.

at the end of the tail. Most living birds have pneumatic cervical and thoracic vertebrae¹⁵, and some, such as the turkey (*Meleagris*) and the ostrich (*Struthio*), exhibit pneumatization of the synsacral and free caudal vertebrae as well.

Identification of pneumatic foramina in *Archaeopteryx* links the evidence for axial pneumaticity seen in non-avian theropods with the pneumatic features of Aves. According to present theories of bird origins¹⁶, this phylogenetic continuity shows the homology of these structures. Although pneumatic postcranial bones are known in advanced pterosaurs¹⁷ and sauropod dinosaurs¹⁸, phylogenetic hypotheses^{16,19,20} indicate that these incidences of postcranial pneumatic invasion occurred separately from that in theropods. Under these phylogenetic schemes, the distribution of vertebral pneumatization within Theropoda supports the proposed relationship of birds to theropods. This character distribution also shows that although axial pneumaticity may lighten the skeleton, its evolution cannot be considered to be an adaptation for flight *per se*.

A flow-through lung is unique to birds among extant vertebrates. This remarkable lung is distinctive for its associated system of air sacs, diverticula of which often invade the postcranial skeleton. The presence of pneumatic foramina in the vertebrae of many non-avian and avian theropods indicates that at least one of the components of the avian air-sac lung system, the cervical air sacs, was already in place in non-avian theropods. In chickens (*Gallus*), the synsacral vertebrae are pneumatized by the abdominal air sacs²¹, whereas in turkeys the cervical air sac extends all the way to the free coccygeal vertebrae²¹. It is impossible at present to determine which of these two air sacs is responsible for invading this part of the vertebral column in non-bird theropods with pneumatic sacral vertebrae, but it is possible that abdominal air sacs were present in these taxa. Clavicular air sacs are only positively known in Aves, although a large pneumatic foramen occurs on a single enantiornithine humerus (L. Chiappe, personal communication). Certain air sacs rarely (abdominal) or possibly never (thoracic) invade the skeleton in living birds¹⁰, and their presence therefore remains enigmatic in stem avialans and more primitive theropods.

The evidence for components of the avian respiratory apparatus in non-avian theropods raises questions about a claim that non-avian theropods and Avialae have different accessory ventilatory mechanisms²², the former having a hepatic-piston pump and the latter an air-sac system. The proposed presence of a hepatic-piston pump in non-avian theropods is based on a preservational artefact (P. Currie, personal communication). Furthermore, the predictions of this hypothesis regarding the distribution of extreme opistho-

puby among non-avian and basal avian theropods has not been supported by some fossil discoveries^{14,23}.

Received 29 January; accepted 7 July 1998.

- Owen, R. On the *Archaeopteryx* of von Meyer, with a description of a new long tailed species from the lithographic stone of Solenhofen [sic]. *Phil. Trans.* **153**, 33–47 (1864).
- Vogt, C. *Archaeopteryx macrura*, an intermediate form between birds and reptiles. *Ibis* **4**, 434–456 (1890).
- Lambrecht, L. *Handbuch der Palaeornithologie* 1024 (Gebrüder Borntraeger, Berlin, 1933).
- De Beer, Sir G. *Archaeopteryx lithographica* (Brit. Mus. Nat. Hist., London, 1954).
- Bühler, P. in *Papers in Avian Paleontology Honoring Pierce Brodkorb* Science Ser. 36 (ed. Campbell, K. E. Jr) (Los Angeles County Mus., Los Angeles, 1992).
- Martin, L. D. in *Origins of the Higher Groups of Tetrapods, Controversy and Consensus* (eds Schultze, H.-P. & Trueb, L.) 485–540 (Cornell Univ. Press, Ithaca, 1991).
- Müller, B. The air-sacs of the pigeon. *Smithson. Misc. Coll.* **50**, 365–414 (1908).
- Wellnhofer, P. Das fünfte Skelettexemplar von *Archaeopteryx*. *Palaeontograph. A* **147**, 169–216 (1974).
- Ostrom, J. H. *Archaeopteryx* and the origins of birds. *Biol. J. Linn. Soc.* **8**, 91–182 (1976).
- King, A. S. Structural and functional aspects of the avian lungs and air sacs. *Int. Rev. Gen. Exp. Zool.* **2**, 171–267 (1966).
- Britt, B. B. *Pneumatic Postcranial Bones in Dinosaurs and Other Archosaurs*. Thesis, Univ. Calgary (1993).
- Hogg, D. A. The development of pneumatization in the postcranial skeleton of the domestic fowl. *J. Anat.* **139**, 105–113 (1984).
- Britt, B. B. in *Encyclopedia of Dinosaurs* (eds Currie, P. C. & Padian, K.) 590–598 (Academic, San Diego, 1997).
- Forster, C. A., Sampson, S. D., Chiappe, L. M. & Krause, D. W. The theropod ancestry of birds: new evidence from the Late Cretaceous of Madagascar. *Science* **279**, 1915–1919 (1998).
- Boas, J. E. V. Studien über den Hals des Vogel. *Kgl. Danske Vidensk. Selsk. Skrifter* **9**, 101–222 (1929).
- Gauthier, J. Saurischian monophyly and the origin of birds. *California Acad. Sci. Mem.* **8**, 1–55 (1986).
- Wild, R. Die Flugsaurier (Reptilia, Pterosauria) aus der Oberen Trias von Cene bei Bergamo, Italien. *Boll. Soc. Paleont. Ital.* **17**, 176–256 (1978).
- McIntosh, J. S. in *The Dinosauria* (eds Weishampel, D. B., Dodson, P. & Osmólska, H.) 345–401 (Univ. California Press, Berkeley, 1990).
- Sereno, P. C. Basal archosaurs: phylogenetic relationships and functional implications. *Soc. Vertebr. Paleont. Mem.* **22**, 1–53 (1991).
- Juul, L. The phylogeny of basal archosaurs. *Palaeont. Afr.* **31**, 1–38 (1994).
- King, A. S. in *The Anatomy of the Domestic Animals* (ed. Getty, R.) 1883–1918 (W. B. Saunders, Philadelphia, 1975).
- Ruben, J. A., Jones, T. D., Geist, N. R. & Jaap Hillenius, W. Lung structure and ventilation in theropods and early birds. *Science* **278**, 1267–1270 (1997).
- Norell, M. A. & Makovicky, P. J. Important features of the dromaeosaur skeleton: information from a new specimen. *Am. Mus. Nov.* **3215**, 1–28 (1997).
- Bonde, N. in *The Continental Jurassic* (ed. Morales, M.) 193–199 (Mus. North. Arizona Bull. 60, Flagstaff, 1996).

Acknowledgements. We thank the curatorial staffs of the American Museum of Natural History, British Museum of Natural History, Humboldt Museum für Naturkunde, Jura-Museum Eichstätt and Yale Peabody Museum for access to their specimens; and L. Barretti, M. Ellison and A. Gishlick for assistance with the production of the figures. This work was supported by the Royal Tyrrell Museum and the University of Calgary (B.B.B.), the Danish Research Academy and the AMNH (P.J.M.), an NSF grant and a travel grant from the Geraldine Lindsay Fund of the California Academy of Sciences (to J.G.), and a travel grant from the Geological Institute, Copenhagen University (to N.B.).

Correspondence and requests for materials should be addressed to P.J.M. (e-mail: pmako@amnh.org).

Object-based attention in the primary visual cortex of the macaque monkey

Pieter R. Roelfsema, Victor A. F. Lamme & Henk Spekreijse

Graduate School Neurosciences Amsterdam, Laboratory of Medical Physics, and the Netherlands Ophthalmic Research Institute (KNAW), Department of Visual System Analysis, Academic Medical Center (UvA), PO Box 12141, 1100 AC Amsterdam, The Netherlands

Typical natural visual scenes contain many objects, which need to be segregated from each other and from the background. Present theories subdivide the processes responsible for this segregation into a pre-attentive and attentive system^{1,2}. The pre-attentive system segregates image regions that 'pop out' rapidly and in parallel across the visual field. In the primary visual cortex, responses to pre-attentively selected image regions are enhanced^{3–5}. When objects do not segregate automatically from the rest of the image, the time-consuming attentive system is recruited. Here we investigate whether attentive selection is also associated with a modulation of firing rates in area V1 of the brain in monkeys trained to perform a curve-tracing task^{6,7}. Neuronal responses to the various segments of a target curve

were simultaneously enhanced relative to responses evoked by a distractor curve, even if the two curves crossed each other. This indicates that object-based attention is associated with a response enhancement at the earliest level of the visual cortical processing hierarchy.

Visual attention has been suggested to act like a spotlight⁸ or zoom lens⁹ that is directed successively to image regions of interest. Neurons in extrastriate areas exhibit enhanced firing rates if attention is directed to the image location to which they respond^{10–15}. Whether attentive rate modulation also occurs in the primary visual cortex is more controversial. Motter¹⁵ demonstrated that visual attention is associated with a substantial enhancement of firing rates in area V1, but other studies have reported weaker¹⁶ or marginal effects^{12,17,18}. Attentive selection cannot be based entirely on the spatial position of a spotlight, however, because it can also be directed to one of two overlapping objects^{19,20}. To select all features of an object in these situations, attention should be guided by perceptual grouping criteria such as connectedness and collinearity^{21,22}. Because neurons in area V1 are sensitive for collinear arrangements of contour segments^{23–25}, we hypothesized that early visual areas might contribute to attentive selection. Activity in early visual areas might also enhance the spatial resolution of the selection process, which is important if components of different objects are close together.

Two monkeys were trained to perform a curve-tracing task^{6,7}. In each trial the animals had to fixate a small circle in the middle of a computer screen (Fig. 1a). After 300 ms of fixation, two red circles and two curves appeared on the screen. One of the circles was connected to the fixation point through a curve, and served as target. The second circle was a distractor, and was not connected to the fixation point. After an additional delay of 600 ms the fixation point was extinguished, and an eye movement had to be made to the target circle. When the monkeys had learned the task, 40–50 multiunit electrodes were chronically implanted in the primary visual cortex.

Figure 1b, c illustrates the location of the receptive field of a recording site in area V1 relative to the visual stimuli. For the first stimulus (Fig. 1b) the receptive field was on the target curve, which connected the fixation point to the target circle. A small change in the stimulus close to the fixation point switched the identity of target and distractor (Fig. 1c) and, accordingly, the receptive field was on the distractor curve. The activity of the neurons was modulated by the difference between stimuli, although the location of the difference was far from the receptive field. The firing rate was highest when the receptive field was on the target curve (Fig. 1d). An analysis of the distributions of firing rates in single trials indicated that this rate enhancement was highly significant (*U*-test, $P < 10^{-6}$) (Fig. 1e).

A comparable enhancement of responses to the target curve was observed for the majority of recording sites in area V1. For all sites ($N = 45$), a modulation index was computed, which was defined as the ratio between rate enhancement (or reduction) and the average firing rate (see Methods). The distribution of modulation indices was shifted to positive values ($P < 0.0005$, sign-test), and had a median of 0.27 (Fig. 1f). This indicates that most cells in area V1 fire more action potentials if their receptive field is on the curve from the fixation point to the target.

To determine the latency of enhancement, responses at all recording sites with a significant rate enhancement ($P < 0.05$ and a positive modulation index, $N = 27$) were normalized to the peak response and averaged (see Methods) (Fig. 1g). The latency was defined as the first of three successive 10-ms bins with a significant difference between the two conditions ($P < 0.05$, paired *t*-test). At the population level, the latency was 235 ms, but in some individual cases it occurred slightly earlier (Fig. 1d). The visual response had a latency of 35 ms and preceded the rate enhancement by 200 ms. These results indicate that, whereas the early activity of a neuron in

area V1 signals features of the contour segment inside the classical receptive field, later response components are influenced from remote locations. By the time of the response enhancement, transients evoked by the appearance of the stimulus have vanished from area V1 (Fig. 1g), which justifies computation of the strength of modulation during the ensuing episode of sustained firing.

The monkeys maintained visual fixation within a square window

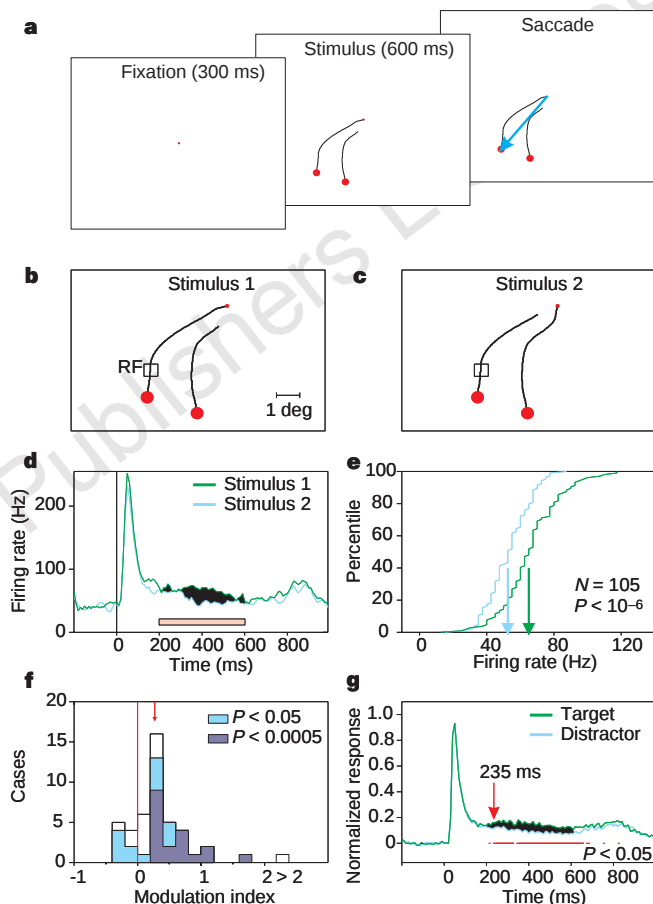


Figure 1 Response enhancement in area V1 during curve tracing. **a**, Sequence of images on the computer monitor during a single trial. Fixation had to be maintained until the fixation point disappeared, 600 ms after stimulus presentation. **b**, **c**, The rectangle shows the location of the receptive field (RF) of a group of neurons in area V1 relative to the stimuli. In **b** the receptive field was on the path from the fixation point to the target and in **c** it was not. **d**, Responses to the two stimuli. Green curve, response to stimulus 1. Blue curve, response to stimulus 2. The black area emphasizes the difference between the responses to the two stimuli during the computational window. Pink bar, computational window used for the analysis of firing rates in single trials. **e**, Distributions of firing rates in a window from 200 to 600 ms after stimulus onset, evoked by stimulus 1 (receptive field on path, green curve) and stimulus 2 (receptive field not on path, blue curve). The monkey performed 105 trials with both stimuli. The arrows indicate the medians of the two distributions that were significantly different (*U*-test, $P < 10^{-6}$). **f**, Distribution of the modulation index of all recording sites ($N = 45$). A positive modulation index indicates an enhanced response to the target curve (see Methods). Cases for which the difference between conditions was significant ($P < 0.05$, *U*-test) are shown in light blue, and highly significant cases are shown in dark blue ($P < 0.0005$). The arrow indicates the median (0.27). **g**, The latency of rate enhancement was estimated by averaging normalized responses at recording sites with a significantly positive modulation index ($P < 0.05$). Green curve, average response when the receptive field was on the target curve. Blue curve, average response when the receptive field was on the distracting curve. Red bars indicate 10-ms bins in which the difference between responses reached statistical significance (paired *t*-test, $P < 0.05$). Arrow, latency of response enhancement.

that was less than or equal to $1^\circ \times 1^\circ$. Nevertheless, microsaccades (small eye movements around the fixation point) occurred. A systematic difference in the pattern of microsaccades between stimuli could cause a modulation of firing rates, because it would be associated with a systematic shift of curves relative to the receptive fields. A stratification procedure was used to exclude this possibility. First, trials in which a microsaccade occurred within a window from 300 to 460 ms after stimulus onset were excluded from analysis (Fig. 2a). For the remaining trials of both stimulus conditions, distributions of eye positions were computed (Fig. 2b). Subsequently, the number of trials in each eye-position bin ($0.2^\circ \times 0.2^\circ$) was made equal by removing excess trials in one or other condition. After this stratification procedure, firing rates were recalculated in a window from 400 to 500 ms after stimulus onset. This shorter window also excludes contributions of transient responses evoked by microsaccades occurring slightly before 300 ms. Figure 2c illustrates the effect of stratification on the

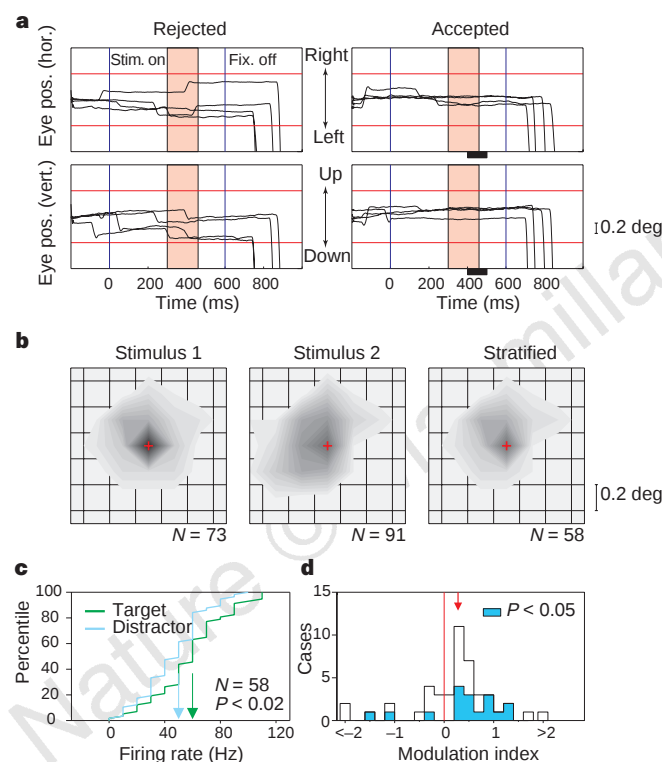


Figure 2 Rate enhancement is not caused by differences in fixation behaviour between stimulus conditions. **a**, Recordings of eye position during curve tracing. Upper panels, horizontal eye position. Lower panels, vertical eye position. Eye positions were determined for each trial in a window from 300 to 460 ms after stimulus onset (pink region). Eye position is only well defined for trials without a microsaccade in this window (right). Trials with a microsaccade in this window (left) were excluded from analysis. Red lines indicate the borders of the fixation window. Black rectangles below the abscissa indicate the time window used for computation of the firing rate distributions. **b**, Distribution of eye positions ($0.2^\circ \times 0.2^\circ$ bins) for the experiment shown in Fig. 1. Left, eye positions in trials with stimulus 1 (N = 73, after removal of trials with a microsaccade). Middle, eye positions for stimulus 2 (N = 91). Right, the stratified distribution of eye positions (N = 58) was obtained by removing excess trials in every bin. Red cross, location of fixation point. **c**, Distributions of firing rates at the same recording site as in Fig. 1 for the subset of trials obtained by stratification. Green curve, firing rates when the receptive field was on the target curve. Blue curve, firing rates when the receptive field was on the distractor curve. The arrows indicate the medians of the two distributions, which differed significantly (U -test, $P < 0.02$). **d**, Distribution of modulation indices after stratification (N = 45). Significant cases are shown in blue ($P < 0.05$, U -test). The arrow indicates the median modulation index (0.28).

distribution of firing rates at the recording site of Fig. 1. Exclusion of trials and the use of a shorter temporal window reduced the significance of the difference between conditions, but the magnitude of the modulation index was virtually unaltered, for the illustrated recording site ($P < 0.02$, U -test) and also at the population level (median = 0.28; $P < 0.005$, sign test) (Fig. 2d). Thus, rate enhancement is not caused by a systematic difference in visual fixation between the two stimulus conditions.

Rate enhancement was not restricted to neuronal responses to a specific segment of the target curve. Figure 3 illustrates an experiment in which simultaneous recordings were obtained from three groups of neurons, which responded to different segments of a single curve. When the respective receptive fields were on the target curve, responses at all recording sites were enhanced simultaneously. This demonstrates that the representation of the entire target curve 'lights up' in area V1 during this task.

Unlike response modulations observed in previous studies^{3–5,23–25}, the rate enhancement during curve tracing cannot be attributed to a pre-attentive mechanism, because the monkeys had to select one of two equally salient curves. Results from psychophysical studies are also inconsistent with pre-attentive processing in curve-tracing tasks, because reaction times exhibit an approximately linear dependence on the length of curve that should be traced^{6,7}. We therefore conjecture that rate enhancement during curve tracing is due to attentive selection and that the latency of enhancement provides a measure for the time required to direct attention to the curve segment inside the receptive field.

Treisman and co-workers (for example, ref. 1) have suggested that focal attention needs to be directed to a visual object to integrate its various attributes into a coherent object representation. The curve-tracing task would have been solved as soon as visual attention had grouped contour segments into a coherent representation of the target curve, because this would identify the correct eye-movement target. However, in many of the stimuli used, it was virtually impossible for a circular spotlight to enclose the entire target curve while avoiding the distractor curve (Figs 1, 3). Nevertheless, responses to the various segments of the target curve were enhanced simultaneously, which suggests that object-based attention^{19,22} is involved in curve tracing.

The case for the involvement of object-based attention would be strengthened if a selective enhancement of responses to the target curve were also to occur if the two curves were spatially overlapping. We therefore investigated activity in V1 by using a slightly modified

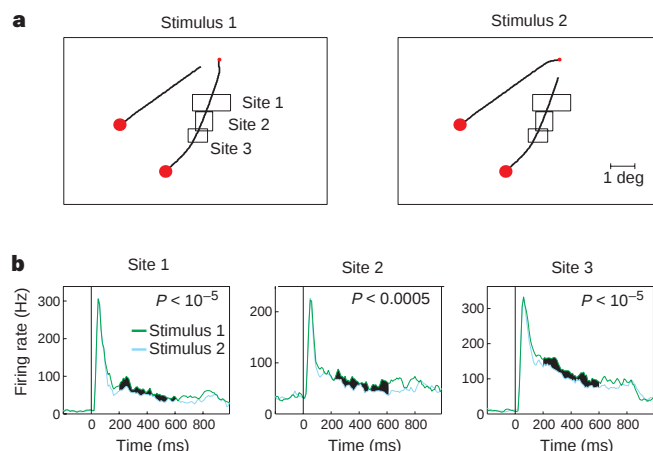


Figure 3 Simultaneous enhancement of responses to various segments of a single curve. **a**, Rectangles show the location of receptive fields at three recording sites from which simultaneous recordings were obtained. Receptive fields were all on the target curve for stimulus 1 (left), and on the distracting curve for stimulus 2 (right). **b**, Responses at the recording sites were simultaneously enhanced when the receptive fields were on the target curve.

task in which the target curve could cross the distractor curve. An additional location of stimulus variation, the 'switch-box', was introduced (inset in Fig. 4a, b). The two curves could intersect each other in the switch-box or stay separate. Figure 4 illustrates the location of the receptive fields of five simultaneously recorded groups of V1 neurons relative to the stimuli. When the two curves were non-intersecting (Fig. 4a), firing rates were enhanced significantly (yellow receptive fields; $P < 0.05$, difference between stimulus I and II) at recording sites that had their receptive field on the target curve, which is in accordance with the results of Figs 1 and 3. Notably, even if the two curves crossed each other, responses to the various segments of the target curve were simultaneously enhanced (Fig. 4b; yellow receptive fields). This enhancement was significant at four of the five recording sites ($P < 0.05$, difference between stimulus III and IV).

Cells at recording site 2 had their receptive field on the proximal curve segment, between the fixation point and the location at which an intersection could occur. Response enhancement at site 2 occurred for stimulus II and IV, although these stimuli were associated with different saccade targets (Fig. 4a–c). This indicates

that rate enhancement does not depend solely on the location of fixation point and the planned eye movement, but rather on the position of the entire target curve. The population response of all recording sites with a receptive field on the proximal segment ($N = 15$) corroborates this result, and is shown in Fig. 4e. On average, responses were enhanced if the receptive field was on the target curve, for stimuli with (modulation index = 0.35; paired t -test, $P < 0.005$) and without (modulation index = 0.18; paired t -test, $P < 0.005$) an intersection.

The receptive field of neurons at recording site 3 was on the distal curve segment, between the switch-box and the circular targets. Response enhancement at this site occurred for stimulus I and IV (Fig. 4a, b, d). Thus, if there was no intersection, the response at recording site 3 was strongest if the fixation point was connected to the upper proximal segment, but this dependence was reversed in the presence of an intersection. This indicates that the response enhancement is not a simple sensory effect, caused by the presence of an additional contour segment at a specific location in the visual field. The population response of all recording sites with a receptive field on the distal segment ($N = 23$) confirms these results

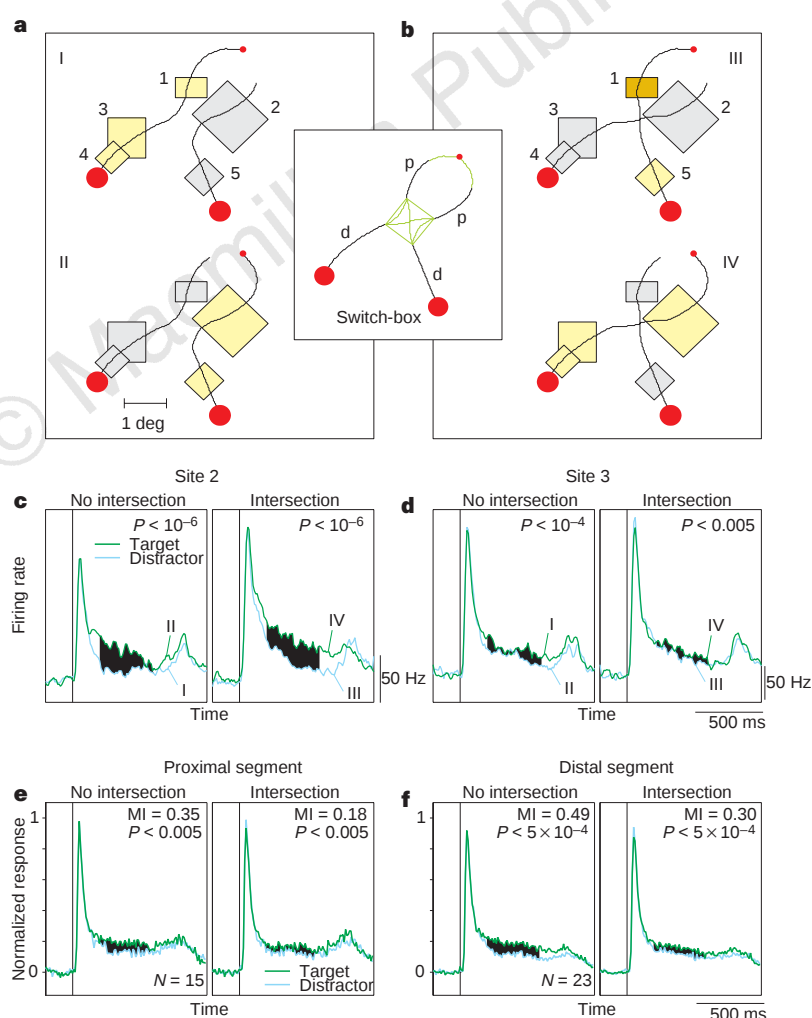


Figure 4 Response enhancement to stimuli in which the target and distractor curve could intersect each other. **a, b**, Complementary stimuli with (**b**) or without (**a**) an intersection. Rectangles show the receptive fields at five recording sites. Receptive fields in bright yellow are from recording sites at which responses to the target curve were significantly enhanced ($P < 0.05$; comparison between stimulus I and II, and between III and IV). The response enhancement at recording site 1 to stimulus III did not reach statistical significance (darkish yellow). Inset, design of the stimuli. Green contour segments varied between stimuli, black

contour segments were identical. The green square indicates the switch-box. p, proximal curve segment. d, distal curve segment. **c, d**, Responses to complementary stimuli at recording site 2 (**c**) and recording site 3 (**d**). Responses to the target curve are shown in green, and responses to the distractor curve in blue. **e, f**, Responses pooled across all recording sites with a receptive field on the proximal curve segment (**e**), and distal curve segment (**f**). N , number of recording sites. The modulation indices (MI) refer to the medians of the MI distributions across recording sites.

(Fig. 4f). Responses were stronger, on average, if the receptive field was on the target curve, both for stimuli without an intersection (modulation index = 0.49; paired *t*-test, $P < 5 \times 10^{-4}$), as well as for stimuli with an intersection (modulation index = 0.30; paired *t*-test, $P < 5 \times 10^{-4}$). The population responses of Fig. 4e, f provide additional evidence against the possibility that response modulation is caused by a purely sensory effect related to the difference between stimuli close to the fixation point. The proximal curve segment was closest to the fixation point, but modulation was strongest for neurons with a receptive field on the distal curve segment.

In conclusion, neurons that respond to segments of an entire target curve simultaneously exhibit an enhanced firing rate, even if this curve crosses a second, irrelevant curve. This strongly suggests that rate modulations in area V1 provide a correlate of object-based attention. To label responses to one of the curves selectively, the distribution of the rate enhancement should depend on perceptual grouping criteria such as collinearity and connectedness. Horizontal connections in area V1 predominantly interconnect neurons that prefer collinear line elements and have closely spaced receptive fields^{26,27}. These connections might be involved in the propagation of attentive rate modulations to neurons that respond to the various segments of a single curve. The high degree of positional specificity of V1 neurons (small receptive fields) might be essential for the selectivity of such a process in situations in which different curves come into close proximity. □

Methods

Curve-tracing task. Experiments were performed with two macaque monkeys. Monkeys were seated in a primate chair with the head restrained at a distance of 0.75 m from a computer monitor (resolution 1024×768 , frame rate 70 Hz). The eye position was measured by using a double magnetic induction technique²⁸. A trial was started as soon as the monkey's eye position was within a $1^\circ \times 1^\circ$ (or $0.8^\circ \times 0.8^\circ$) square window centred on the fixation point (0.2° diameter). After an interval of 300 ms, the targets and curves were shown (Fig. 1a), but the monkey had to maintain fixation. The target and fixation point were red, and the curves white (luminance 85 cd m^{-2}) on a black background (luminance 1.5 cd m^{-2}). After an additional 600 ms the fixation point was extinguished, and the monkey made an eye movement to one of the targets. Eye movements to the target that was connected to the fixation point were rewarded with apple juice.

Surgical procedures. The monkeys underwent two operations under general anaesthesia. Anaesthesia was induced with ketamine (15 mg kg^{-1} injected intramuscularly) and was maintained after intubation by ventilation with a mixture of 70% N_2O and 30% O_2 , supplemented with 0.8% isoflurane, fentanyl (0.005 mg kg^{-1} intravenously) and midazolam ($0.5 \text{ mg kg}^{-1} \text{ h}^{-1}$ intravenously). In the first operation a head holder was implanted. In addition, a gold-plated copper ring was implanted under the conjunctiva of one eye for the measurement of eye position²⁸. In the second operation 40–50 Teflon-coated platinum–iridium wires (diameter $25 \mu\text{m}$, impedance $0.4\text{--}0.8 \text{ M}\Omega$ at 100 Hz) were implanted chronically in area V1. The tips of the wires were positioned 1–2 mm below the cortical surface. Therefore our sample is relatively devoid of recordings in the upper cortical layers. The animals recovered for at least 21 days before training was resumed and data collection was initiated. All procedures complied with the NIH Guide for Care and Use of Laboratory Animals (National Institute of Health, Bethesda, Maryland), and were approved by the institutional animal care and use committee of the Royal Netherlands Academy of Arts and Sciences.

Recording and data analysis. The eye position was recorded in every trial with a sampling rate between 0.5 and 1 kHz, and microsaccades were detected offline by an automated procedure that convoluted the eye position traces with a step function. For detection of multi-unit activity, signals were amplified, band-pass filtered and fed through a Schmitt trigger. Recordings with a sufficient signal-to-noise ratio were obtained from about 50% of the wires. For these recording sites, receptive field dimensions were determined with an automatic plotting procedure using moving light bars. Receptive field eccentricity ranged from 1° to 6° . Average receptive field size was 0.66 deg^2

(range $0.14\text{--}1.6 \text{ deg}^2$), which is well within the range of V1 receptive field dimensions obtained with single-cell recordings in awake monkeys²⁹, and with conventional multi-unit recordings in anaesthetized monkeys³⁰. The significance of differences in response strength between stimuli was computed from the distributions of firing rates in single trials (Fig. 1e) using the Mann–Whitney *U* test. As a measure of the strength of response enhancement, a modulation index was computed in a computational window from 200 to 600 ms (Fig. 1f), or from 400 to 500 ms (Fig. 2d) after stimulus onset. The modulation index was defined as the difference in response strength normalized to the average: $(T - D)/[(T + D)/2]$, where *T* and *D* are responses to the target and distractor curve, respectively, after subtraction of the spontaneous firing rate. The median modulation index for stimuli without intersections was 0.27, which corresponds to a rate enhancement by 31% ($T/D = 1.31$). To compute the latency of rate enhancement, a population response was computed (Fig. 1g) from the normalized responses at recording sites with a significantly positive modulation index ($P < 0.05$, $N = 27$). To derive the normalization parameters, responses at each recording site (10 ms bins) were first averaged across the two conditions (on-path and off-path). From this average the peak response (R_p) and the spontaneous firing rate (F_s) were determined. Responses to both stimuli were divided by $(R_p - F_s)$, after subtraction of F_s . This normalization procedure would preserve a difference in the peak response between stimuli. Normalized responses were averaged across recording sites, and in each 10 ms bin a paired *t*-test was performed on the distributions of firing rates across recording sites. The latency of response enhancement was identical if responses from all recording sites ($N = 45$) were used for the computation of the population response (including non-significant cases, and cases with a negative modulation index). The population responses to stimuli with the switch box (Fig. 4e, f) were also computed using all cases.

Received 21 May; accepted 9 July 1998.

1. Treisman, A. M. & Gelade, G. A feature-integration theory of attention. *Cogn. Psychol.* **12**, 97–136 (1980).
2. Posner, M. I. & Presti, D. E. Selective attention and cognitive control. *Trends Neurosci.* **10**, 13–17 (1987).
3. Lamme, V. A. F. The neurophysiology of figure-ground segregation in primary visual cortex. *J. Neurosci.* **15**, 1605–1615 (1995).
4. Zipser, K., Lamme, V. A. F. & Schiller, P. H. Contextual modulation in primary visual cortex. *J. Neurosci.* **16**, 7376–7389 (1996).
5. Knierim, J. J. & Van Essen, D. C. Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *J. Neurophysiol.* **67**, 961–980 (1992).
6. Jolicoeur, P., Ullman, S. & MacKay, M. Curve tracing: a possible basic operation in the perception of spatial relations. *Mem. Cognit.* **14**, 129–140 (1986).
7. Pringle, R. & Egeth, H. E. Mental curve tracing with elementary stimuli. *J. Exp. Psychol.: Hum. Percept. Perform.* **14**, 716–728 (1988).
8. Posner, M. I., Snyder, C. R. & Davidson, B. J. Attention and the detection of signals. *J. Exp. Psychol.: Gen.* **109**, 160–174 (1980).
9. Eriksen, C. W. & St. James, J. D. Visual attention within and around the field of focal attention: a zoom lens model. *Percept. Psychophys.* **40**, 225–240 (1986).
10. Mausell, J. H. R. The brain's visual world: representation of visual targets in cerebral cortex. *Science* **270**, 764–769 (1995).
11. Moran, J. & Desimone, R. Selective attention gates visual processing in the extrastriate cortex. *Science* **229**, 782–784 (1985).
12. Luck, S. J., Chelazzi, L., Hillyard, S. A. & Desimone, R. Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *J. Neurophysiol.* **77**, 24–42 (1997).
13. Treue, S. & Mausell, J. H. R. Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature* **382**, 539–541 (1996).
14. Bushnell, M. C., Goldberg, M. E. & Robinson, D. L. Behavioral enhancement of visual responses in monkey cerebral cortex. I. Modulation in posterior parietal cortex related to selective visual attention. *J. Neurophysiol.* **46**, 755–772 (1981).
15. Motter, B. C. Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *J. Neurophysiol.* **70**, 909–919 (1993).
16. Press, W. A. & van Essen, D. C. Attentional modulation of neuronal responses in macaque area V1. *Soc. Neurosci. Abstr.* **23**, 1026 (1997).
17. Wurtz, R. H. & Mohler, C. W. Enhancement of visual responses in monkey striate cortex and frontal eye fields. *J. Neurophysiol.* **39**, 766–772 (1976).
18. Haenny, P. E. & Schiller, P. H. State dependent activity in monkey visual cortex. I. Single cell activity in V1 and V4 on visual tasks. *Exp. Brain Res.* **69**, 225–244 (1988).
19. Duncan, J. Selective attention and the organization of visual information. *J. Exp. Psychol.: Gen.* **113**, 501–517 (1984).
20. Kanwisher, N. & Driver, J. Objects, attributes, and visual attention: which what and where. *Curr. Dir. Psychol. Sci.* **1**, 26–31 (1997).
21. Rock, I. & Palmer, S. The legacy of Gestalt psychology. *Sci. Am.* **263**(6), 48–61 (1990).
22. Kramer, A. F. & Jacobson, A. Perceptual organization and focused attention: the role of objects and proximity in visual processing. *Percept. Psychophys.* **50**, 267–284 (1991).
23. Nelson, J. I. & Frost, B. J. Intracortical facilitation among co-oriented, co-axially aligned simple cells in cat striate cortex. *Exp. Brain Res.* **61**, 54–61 (1985).
24. Kapadia, M. K., Ito, M., Gilbert, C. D. & Westheimer, G. Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in V1 of alert monkeys. *Neuron* **15**, 843–856 (1995).
25. Polat, U., Mizobe, K., Pettet, M. W., Kasamatsu, T. & Norcia, A. M. Collinear stimuli regulate visual responses depending on cell's contrast threshold. *Nature* **391**, 580–584 (1998).

26. Bosking, W. H., Zhang, Y., Schofield, B. & Fitzpatrick, D. Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *J. Neurosci.* **15**, 2112–2127 (1997).
27. Schmidt, K. E., Goebel, R., Löwel, S. & Singer, W. The perceptual grouping criterion of colinearity is reflected by anisotropies of connections in the primary visual cortex. *Eur. J. Neurosci.* **9**, 1083–1089 (1997).
28. Bour, L. J., Van Gisbergen, J. A. M., Bruijns, J. & Ottes, F. P. The double magnetic induction method for measuring eye movement—results in monkey and man. *IEEE Trans. Biomed. Eng.* **31**, 419–427 (1984).
29. Snodderly, D. M. & Gur, M. Organization of striate cortex of alert, trained monkeys (*Macaca fascicularis*): ongoing activity, stimulus selectivity, and widths of receptive field activating regions. *J. Neurophysiol.* **74**, 2100–2125 (1995).
30. Gattass, R., Gross, C. G. & Sandell, J. H. Visual topography of V2 in the macaque. *J. Comp. Neurol.* **201**, 519–539 (1981).

Acknowledgements. We thank J. C. de Feiter and K. Brandsma for technical assistance. The research of P.R.R. and V.A.F.L. was funded by a fellowship from the Royal Netherlands Academy of Arts and Sciences.

Correspondence and requests for materials should be addressed to P.R.R. (e-mail: p.roelfsema@ioi.knaw.nl).

An analgesia circuit activated by cannabinoids

Ian D. Meng*, Barton H. Manning*, William J. Martin† & Howard L. Fields*‡

Departments of *Neurology, †Anatomy and ‡Physiology, and the W. M. Keck Foundation Center for Integrative Neuroscience, University of California, San Francisco, California 94143-0114, USA

Although many anecdotal reports indicate that marijuana and its active constituent, delta-9-tetrahydrocannabinol (delta-9-THC), may reduce pain sensation^{1,2}, studies of humans have produced inconsistent results^{3–6}. In animal studies, the apparent pain-suppressing effects of delta-9-THC and other cannabinoid drugs^{7–12} are confounded by motor deficits^{13,14}. Here we show that a brainstem circuit that contributes to the pain-suppressing effects of morphine¹⁵ is also required for the analgesic effects of cannabinoids. Inactivation of the rostral ventromedial medulla (RVM) prevents the analgesia but not the motor deficits produced

by systemically administered cannabinoids. Furthermore, cannabinoids produce analgesia by modulating RVM neuronal activity in a manner similar to, but pharmacologically dissociable from, that of morphine. We also show that endogenous cannabinoids tonically regulate pain thresholds in part through the modulation of RVM neuronal activity. These results show that analgesia produced by cannabinoids and opioids involves similar brainstem circuitry and that cannabinoids are indeed centrally acting analgesics with a new mechanism of action.

The discoveries of the CB1 and CB2 cannabinoid receptors^{16,17} and of two putative endogenous ligands for the CB1 receptor^{18,19} and the development of a specific CB1-receptor antagonist (SR141716A) (ref. 20) have intensified interest in the function of endogenous cannabinoid systems. The CB2 receptor is located in peripheral tissues, whereas the CB1 receptor is present on neurons throughout the brain, including in several pain-modulating centres²¹. Projections from the RVM to the dorsal horn of the spinal cord are important for the production of analgesia originating from supraspinal sites²². The RVM, which includes the nucleus raphe magnus, the nucleus gigantocellularis pars alpha and the adjacent reticular formation, is essential for systemic opioid-induced analgesia¹⁵, and may also be involved in cannabinoid analgesia, as microinjection of cannabinoid agonists into the RVM suppresses pain-related behaviours²³.

To determine the contribution of the RVM to analgesia produced by a systemically administered cannabinoid, we inactivated the RVM by microinjection of the GABA_A (γ-aminobutyric acid subtype A) receptor agonist muscimol. Rats with physiological saline microinjected into the RVM showed significant increases in tail-flick latencies after intravenous administration of 0.125 mg kg⁻¹ and 0.25 mg kg⁻¹ doses of WIN55,212-2, a cannabinoid receptor agonist (Fig. 1a). In contrast, animals microinjected with muscimol (50 ng) in the RVM before the administration of WIN55,212-2 showed no increase in tail-flick latencies (Fig. 1b). Muscimol microinjection also produced significant hyperalgesia in animals treated with systemic vehicle as compared with control rats microinjected with saline (latency 3.29 ± 0.21 versus 4.73 ± 0.09 s; *P* < 0.05). These results indicate that the activity of neurons in the RVM is necessary

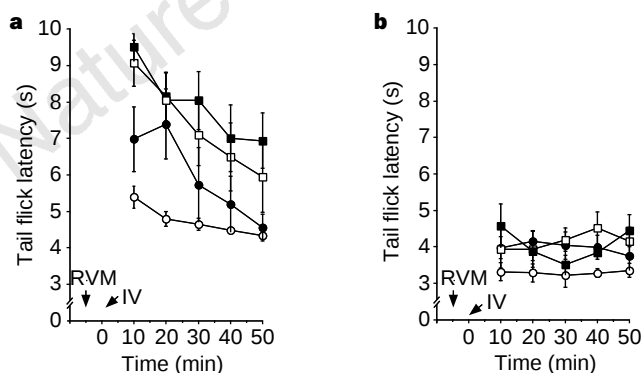


Figure 1 Inactivation of the RVM with microinjection of muscimol prevents the antinociception produced by the cannabinoid agonist WIN55,212-2. We measured tail-flick latencies in groups that received **a**, saline, or **b**, muscimol microinjections into the RVM before intravenous administration of vehicle (open circles); 0.0625 mg kg⁻¹ WIN55,212-2 (filled circles); 0.125 mg kg⁻¹ WIN55,212-2 (open squares); or 0.25 mg kg⁻¹ WIN55,212-2 (filled squares). In saline-injected animals, consistent antinociception was achieved at the 0.125 and 0.25 mg kg⁻¹ doses of WIN55,212-2 (*P* < 0.01; analysis of variance (ANOVA) followed by Fisher's least-significant difference (LSD) test). Muscimol-injected animals showed no difference in tail-flick latencies at any of the WIN55,212-2 doses (*P* > 0.05, ANOVA and Fisher's LSD test). Each value is the mean ± s.e.m.; tail-flick latencies represent the average of three trials; arrows indicate RVM and intravenous (IV) drug administration; *n* = 6 rats per group.

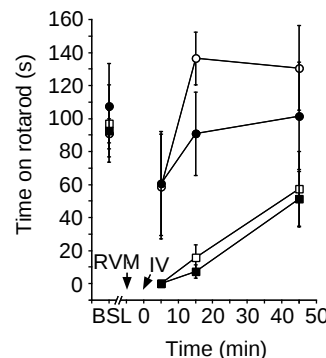


Figure 2 Cannabinoid-induced motor impairments are not affected by inactivation of the RVM with muscimol. Time on the rotarod treadmill (mean ± s.e.m.) was recorded from each of four groups (RVM microinjection/intravenous injection): open circles, saline/vehicle; filled circles, muscimol/vehicle; open squares, saline/WIN55,212-2; filled squares, muscimol/WIN55,212-2. Groups that received WIN55,212-2 (0.25 mg kg⁻¹) spent significantly less time on the rotarod as compared with vehicle-treated groups (*P* < 0.01; ANOVA and Fisher's LSD test). Injection of saline or muscimol into the RVM did not affect rotarod scores (*P* > 0.05; ANOVA). Baseline times (BSL) on the rotarod for each group before treatments were similar. Arrows indicate times of RVM and intravenous (IV) drug administration; *n* = 5 rats per group.

for systemic cannabinoid-induced antinociception, and that the activity of neurons in the RVM contributes to basal nociceptive thresholds.

At antinociceptive doses, cannabinoids produce motor deficits in rats, including hypomotility and catalepsy^{13,14}. Thus, it is necessary to differentiate between the sensory and motor effects of cannabinoids, especially when using a movement such as the tail flick as an index of antinociception^{10,11}. We quantified the effect of RVM inactivation on cannabinoid-induced motor impairments using a rotarod treadmill. Intravenous administration of WIN55,212-2 (0.25 mg kg⁻¹) severely impaired the ability of animals to stay on the rotarod throughout the 45-min test period (Fig. 2). Microinjection of muscimol (50 ng) into the RVM did not prevent the WIN55,212-2-induced motor impairments (Fig. 2). There was no difference between groups given vehicle that were microinjected with saline or muscimol, and baseline times were similar for all groups (Fig. 2, $P > 0.05$). These results show that the prevention of cannabinoid-induced antinociception observed following RVM inactivation is not due to a reversal of cannabinoid-induced motor deficits.

We used extracellular single-unit recording to determine which type of RVM neuron contributes to cannabinoid-induced antinociception. Three physiological classes of RVM neurons have been described²⁴. 'Off' cells inhibit spinal nociceptive transmission and show a pause in activity just before withdrawal reflexes. Conversely, 'on' cells show a burst of activity just before withdrawal reflexes and have a facilitating effect on pain transmission. The activity of a third group of neurons—neutral cells—is not correlated with withdrawal reflexes. While recording from RVM neurons, we intravenously administered anaesthetized rats with, successively, WIN55,212-2 (0.125 mg kg⁻¹), the cannabinoid receptor antagonist SR141716A

(0.5 mg kg⁻¹), morphine sulphate (2.0 mg kg⁻¹) and naloxone (0.2 mg kg⁻¹). The effect of WIN55,212-2 on 'on' and 'off' cell activity was similar to that of morphine (Fig. 3a, b). Administration of either WIN55,212-2 or morphine eliminated the 'off' cell pause, the 'on' cell burst, and the tail-flick reflex in all animals tested (six 'off' cells and five 'on' cells). Quantitative analysis of 'off' cell activity during the tail-flick period showed an elimination of the 'off' cell pause (that is, an increase in firing to baseline levels) after WIN55,212-2 administration, with activity increasing from an average rate of 4.4 ± 1.8 spikes s⁻¹ to 24.8 ± 9.7 spikes s⁻¹ ($P < 0.01$). The 'off' cell pause returned (7.3 ± 5.3 spikes s⁻¹) after subsequent administration of SR141716A. Morphine given after SR141716A eliminated the 'off' cell pause in a naloxone-reversible manner (25 ± 12.1 and 3.0 ± 2.3 spikes s⁻¹, respectively). Overall, there was no change in baseline activity of 'off' cells after WIN55,212-2 administration (from 19.1 ± 10.7 to 25.6 ± 9.6 spikes s⁻¹).

The 'on' cell burst during the tail-flick period was reduced to baseline firing levels after the administration of WIN55,212-2 (18.2 ± 6.6 to 1.7 ± 1.7 spikes s⁻¹), and returned to normal following administration of SR141716A (19.5 ± 9.7 spikes s⁻¹; $P < 0.01$). Administration of morphine after SR141716A reduced the 'on' cell burst to levels seen after WIN55,212-2 and was naloxone-reversible (2.0 ± 2.0 and 17.3 ± 9.8 spikes s⁻¹, respectively). There was no change in the baseline activity of 'on' cells after administration of WIN55,212-2 (from 2.6 ± 1.7 to 1.6 ± 1.5 spikes s⁻¹). Of the five neutral cells tested, the activity of three was unaffected, the activity of one decreased, and the activity of one increased after WIN55,212-2 was administered ($>30\%$ change). As 'off' cells project to the spinal cord dorsal horn²⁵ and are the only neurons activated by WIN55,212-2, activation of 'off' cells probably contributes to cannabinoid analgesia by inhibiting dorsal horn nociceptive neurons.

Direct inhibition of 'on' cells by μ -opioids disinhibits 'off' cells²⁶. Furthermore the release of endogenous opioids in the RVM mediates both the inhibition of 'on' cells and the antinociception seen after activation of neurons in the midbrain periaqueductal grey^{27,28}. Systemically administered cannabinoids could inhibit 'on' cells and activate 'off' cells through a similar circuit to that involving endogenous opioids. We tested the involvement of a direct opioid link in the cannabinoid-induced changes in RVM neuronal activity

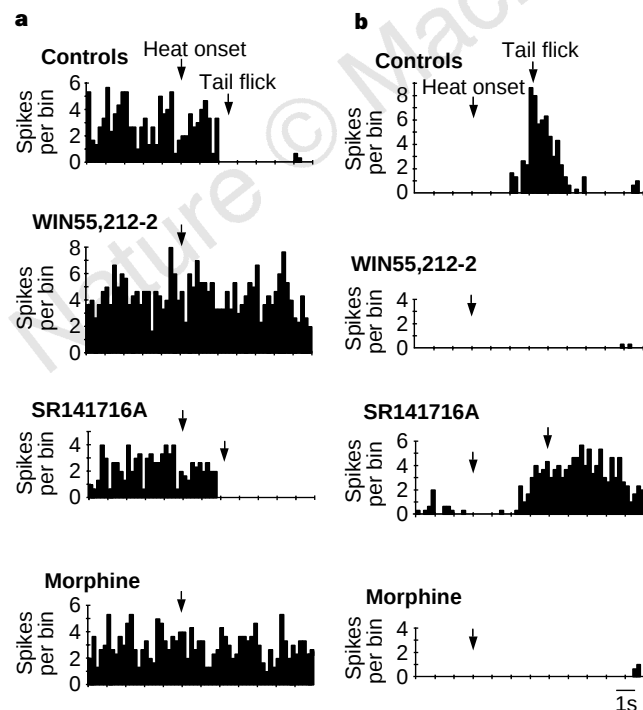


Figure 3 Systemic cannabinoids reduce the 'off' cell pause and 'on' cell burst of RVM neurons, and eliminate the tail-flick reflex in a manner similar to morphine. **a**, An example of 'off' cell activity during control tail flicks and after successive intravenous administrations of WIN55,212-2, SR141716A, and morphine sulphate. **b**, Activity of an 'on' cell during control tail flicks and after successive administrations of WIN55,212-2, SR141716A, and morphine sulphate. Arrows represent the initiation of the thermal stimulus and tail flick. Each histogram represents the average of three trials; bin width, 200 ms.

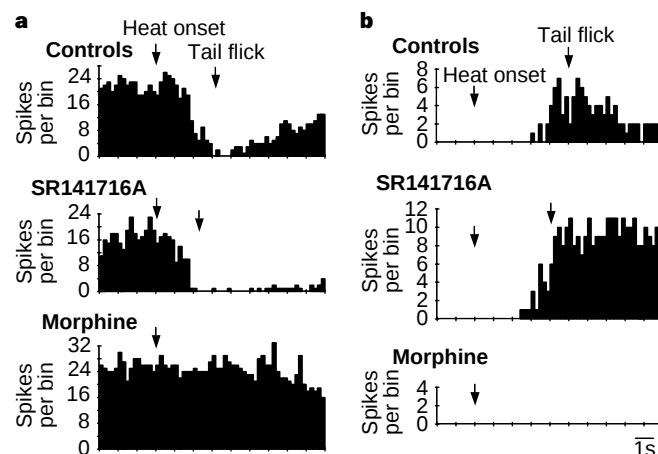


Figure 4 The cannabinoid receptor antagonist SR141716A decreases 'off' cell activity, increases 'on' cell activity and produces hyperalgesia. **a**, Activity of an 'off' cell during control tail flicks, and following successive administrations of SR141716A and morphine sulphate. **b**, 'On' cell activity during control tail flicks, and following successive administrations of SR141716A and morphine sulphate. Arrows represent the initiation of the thermal stimulus and tail flick. Each histogram represents the average of three trials; bin width, 200 ms.

by administering naloxone (0.2 mg kg^{-1} , i.v.) after WIN55,212-2 (0.125 mg kg^{-1} , i.v.) while recording from RVM neurons. Naloxone did not affect the elimination of the 'on' cell burst ($n = 5$) or the 'off' cell pause ($n = 5$) produced by WIN55,212-2 (data not shown), indicating that endogenous opioids are not required for the cannabinoid-induced changes in RVM neuronal activity.

In the mouse, intrathecal administration of SR141716A results in hyperalgesia in the hot-plate test²⁹, indicating that tonically released endogenous cannabinoids modulate nociceptive thresholds. Activity of a precursor enzyme for the synthesis of the endogenous cannabinoid ligand anandamide is highest in the brainstem³⁰. To examine whether a tonically released endogenous cannabinoid controls RVM neurons, we injected SR141716A (0.5 mg kg^{-1}) intravenously, followed by morphine (2.0 mg kg^{-1}) and naloxone (0.2 mg kg^{-1}), while recording from 'on' cells ($n = 5$) and 'off' cells ($n = 5$). Following SR141716A administration, tail-flick latencies decreased from 4.41 ± 0.30 to $3.35 \pm 0.31 \text{ s}$ ($P < 0.01$, $n = 10$). Furthermore, reduction in the 'off' cell tail-flick-related activity was enhanced by more than 30% in all five neurons (for example, Fig. 4a). Baseline activity of 'off' cells also decreased in four out of five neurons ($>30\%$). 'On' cell baseline activity increased in one out of five neurons after administration of SR141716A, and two out of five 'on' cells showed an increase in tail-flick-related activity ($>30\%$) after SR141716A was given (Fig. 4b). Although it is possible that SR141716A acts as an inverse agonist at the CB1 receptor, these results indicate that tonic release of endogenous cannabinoids increases 'off' cell and decreases 'on' cell activity, thereby elevating nociceptive thresholds.

These results show that cannabinoids produce analgesia through activation of a brainstem circuit that is required for opioid analgesia. As with the opioids, activation of 'off' cells in the RVM contributes to cannabinoid-induced antinociception. However, the modulation of 'on' and 'off' cell activity by cannabinoids does not require an endogenous opioid receptor ligand. Furthermore, endogenous cannabinoids seem to be tonically released and to control basal nociceptive thresholds through the modulation of RVM neuronal activity. Given their unique side-effect profile (for example, cannabinoids increase appetite, whereas opioids can cause nausea and respiratory depression), cannabinoids may be useful in improving the treatment of pain. □

Methods

Microinjections. Male Sprague–Dawley rats (275–350 g) were implanted with chronic jugular catheters, and 26-g guide cannulae were positioned 2 mm above the RVM (AP -2.1 (interaural), ML 0, DV 8.5)³¹ and cemented to the skull with dental acrylic. Internal cannulae (33-gauge) that protruded 2 mm beyond the tip of the guide cannula were inserted into the guides for injections. Injector cannulae were connected to 100- μl Hamilton syringes through PE10 tubing backfilled with water. Microinjections of drugs into the RVM (500 nl) were made with an infusion pump over a period of 5 min. All injection sites were histologically verified to be within the RVM.

Rats were handled and adapted to a plexiglas chamber with a glass bottom for 3 days before testing. Tail-flick latencies were measured in awake, unrestrained rats using radiant heat from a modified Hargreaves thermal stimulator. On the test day, intravenous injection of drugs was given 5 min following RVM microinjections. Tail-flick measurements were taken every 2 min for 60 min after administration of intravenous drugs. In rats tested on the rotarod treadmill, initial training determined the ability of each rat to walk for at least 60 s at 10 rotations per min. In all other trials, the rotarod started at 6 rotations per min and accelerated by 2 rotations every min. A pretreatment score was recorded, and was followed by RVM microinjection and, 5 min later, by intravenous drug administration. Rats were tested for their ability to stay on the rotarod at 5, 15 and 45 min following intravenous drugs.

Electrophysiology. Male Sprague–Dawley rats (300–475 g) were anaesthetized with sodium pentobarbital. A jugular vein was catheterized for constant infusion of anaesthetic (sodium methohexital, $15\text{--}30 \text{ mg kg}^{-1} \text{ h}^{-1}$) and administration of drugs. Single RVM neurons were recorded extracellularly

with tungsten electrodes ($3\text{--}4 \text{ M}\Omega$). 'On' cells showed a burst of activity and 'off' cells a pause just before a tail flick. Neutral cells showed no consistent change in activity related to the tail flick. Tail-flick-related activity was examined every 3 min. After 3 stable control tail flicks, rats were given drugs in successive order every 12–18 min. Baseline neuronal activity for 10 s before the heat onset was compared before and after drug administration. Tail-flick-related neuronal activity was defined as the frequency of firing during a 4-s window around the average control tail-flick latency calculated over all experiments. Only one neuron was tested per animal.

Materials. WIN 55,212-2 was purchased from Research Biochemicals International. SR141716A was a gift from the US National Institute of Drug Abuse. Vehicle consisted of ethanol, Alkamuls EL-620 (Rhône-Poulenc) and 0.9% saline at 1:1:18.

Received 1 May; accepted 7 July 1998.

- Mikuriya, T. H. *Marijuana: Medical Papers 1839–1972*. (Medi-Comp Press, Oakland, California, 1973).
- Dunn, M. & Davis, R. The perceived effects of marijuana on spinal cord injured males. *Paraplegia* **12**, 175 (1974).
- Noyes, R., Brunk, S. F., Avery, D. H. & Canter, A. The analgesic properties of delta-9-tetrahydrocannabinol and codeine. *Clin. Pharmacol. Ther.* **18**, 84–89 (1975).
- Clark, W. C., Janal, M. N., Zeidenberg, P. & Nahas, G. G. Effects of moderate and high doses of marihuana on thermal pain: a sensory decision theory analysis. *J. Clin. Pharmacol.* **21** (Suppl.), 299–310 (1981).
- Libman, E. & Stern, M. H. The effects of delta-9-THC on cutaneous sensitivity and its relation to personality. *Person. Individ. Diff.* **6**, 169–174 (1985).
- Holdcroft, A. et al. Pain relief with oral cannabinoids in familial Mediterranean fever. *Anaesthesia* **52**, 483–488 (1997).
- Buxbaum, D., Sanders-Bush, E. & Efron, D. Analgesic activity of tetrahydrocannabinol (THC) in the rat and mouse. *Fed. Proc.* **28**, 735 (1969).
- Sofia, R. D., Nalepa, S. D., Harakel, J. J. & Vassar, H. B. Anti-edema and analgesic properties of delta-9-tetrahydrocannabinol (THC). *J. Pharmacol. Exp. Ther.* **186**, 646–655 (1973).
- Moss, D. E. & Johnson, R. L. Tonic analgesic effects of delta-9-tetrahydrocannabinol as measured with the formalin test. *Eur. J. Pharmacol.* **61**, 313–315 (1980).
- Lichtman, A. H. & Martin, B. R. Cannabinoid-induced antinociception is mediated by a spinal α -noradrenergic mechanism. *Brain Res.* **559**, 309–314 (1991).
- Martin, W. J., Hohmann, A. G. & Walker, J. M. Suppression of noxious stimulus-evoked activity in the ventral posterolateral nucleus of the thalamus by a cannabinoid agonist: correlation between electrophysiological and antinociceptive effects. *J. Neurosci.* **16**, 6601–6611 (1996).
- Herzberg, U., Eliav, E., Bennett, G. J. & Kopin, I. J. The analgesic effects of R(+)-WIN 55,212-2 mesylate, a high affinity cannabinoid agonist, in a rat model of neuropathic pain. *Neurosci. Lett.* **221**, 157–160 (1997).
- Gough, A. L. & Olley, J. E. Delta-9-tetrahydrocannabinol and the extrapyramidal system. *Psychopharmacol.* **54**, 87–99 (1977).
- Ueki, S. Abnormal behavior induced by delta-9-tetrahydrocannabinol. *Trends Pharmacol. Sci.* **1**, 126–129 (1980).
- Mitchell, J. M., Lowe, D. & Fields, H. L. The contribution of the rostral ventromedial medulla to the antinociceptive effects of systemic morphine in restrained and unrestrained rats. *Neuroscience* **87**, 123–133 (1998).
- Matsuda, L. A., Lolait, S. J., Brownstein, M. J., Young, A. C. & Bonner, T. I. Structure of a cannabinoid receptor and functional expression of the cloned cDNA. *Nature* **346**, 561–564 (1990).
- Munro, S., Thomas, K. L. & Abu-Shaar, M. Molecular characterization of a peripheral receptor for cannabinoids. *Nature* **365**, 61–65 (1993).
- Devane, W. A. et al. Isolation and structure of a brain constituent that binds to the cannabinoid receptor. *Science* **258**, 1946–1949 (1992).
- Stella, N., Schweitzer, P. & Piomelli, D. A second endogenous cannabinoid that modulates long-term potentiation. *Nature* **388**, 773–778 (1997).
- Rinaldi-Carmona, M. et al. SR 14176A, a potent and selective antagonist of the brain cannabinoid receptor. *FEBS Lett.* **350**, 240–244 (1994).
- Tsou, K., Brown, S., Sanudo-Pena, M. C., Mackie, K. & Walker, J. M. Immunohistochemical distribution of cannabinoid CB1 receptors in the rat central nervous system. *Neuroscience* **83**, 393–411 (1998).
- Basbaum, A. I. & Fields, H. L. Endogenous pain control systems: brainstem spinal pathways and endorphin circuitry. *Annu. Rev. Neurosci.* **7**, 309–338 (1984).
- Martin, W. J., Tsou, K. & Walker, J. M. Cannabinoid receptor-mediated inhibition of the rat tail-flick reflex after microinjection into the rostral ventromedial medulla. *Neurosci. Lett.* **242**, 33–36 (1998).
- Fields, H. L., Vanegas, H., Hentall, I. D. & Zorman, G. Evidence that disinhibition of brain stem neurones contributes to morphine analgesia. *Nature* **306**, 684–686 (1983).
- Fields, H. L., Malik, A. & Burstein, R. Dorsal horn projection targets of ON and OFF cells in the rostral ventromedial medulla. *J. Neurophysiol.* **74**, 1742–1759 (1995).
- Heinricher, M. M., Morgan, M. M. & Fields, H. L. Direct and indirect actions of morphine on medullary neurons that modulate nociception. *Neuroscience* **48**, 533–543 (1992).
- Pan, Z. Z. & Fields, H. L. Endogenous opioid-mediated inhibition of putative pain-modulating neurons in rat rostral ventromedial medulla. *Neuroscience* **74**, 855–862 (1996).
- Roychowdhury, S. M. & Fields, H. L. Endogenous opioids acting at a medullary μ -opioid receptor contribute to the behavioral antinociception produced by GABA antagonism in the midbrain periaqueductal gray. *Neuroscience* **74**, 863–872 (1996).
- Richardson, J. D., Aanonsen, L. & Hargreaves, K. M. Hypoactivity of the spinal cannabinoid system results in NMDA-dependent hyperalgesia. *J. Neurosci.* **18** (1998).
- Cadas, H., di Tomaso, E. & Piomelli, D. Occurrence and biosynthesis of endogenous cannabinoid precursor, N-arachidonoyl phosphatidylethanolamine, in rat brain. *J. Neurosci.* **17**, 1226–1242 (1997).
- Paxinos, G. & Watson, C. *The Rat Brain in Stereotaxic Coordinates* (Academic Press, New York, 1986).

Acknowledgements. This work was supported by the National Institute of Drug Abuse, the UCSF Center for the Neurobiology of Addiction and the Canadian MRC.

Correspondence and requests for materials should be addressed to I.D.M. (e-mail: idmeng@phy.ucsf.edu).

Paired-spike interactions and synaptic efficacy of retinal inputs to the thalamus

W. Martin Usrey, John B. Reppas & R. Clay Reid

Department of Neurobiology, Harvard Medical School, 220 Longwood Avenue, Boston, Massachusetts, 02115, USA

In many neural systems studied *in vitro*, the timing of afferent impulses affects the strength of postsynaptic potentials^{1,2}. The influence of afferent timing on postsynaptic firing *in vivo* has received less attention. Here we study the importance of afferent spike timing *in vivo* by recording simultaneously from ganglion cells in the retina and their targets in the lateral geniculate nucleus of the thalamus. When two spikes from a single ganglion-cell axon arrive within 30 milliseconds of each other, the second spike is much more likely than the first to produce a geniculate spike, an effect we call paired-spike enhancement. Furthermore, simultaneous recordings from a ganglion cell and two thalamic targets indicate that paired-spike enhancement increases the frequency of synchronous thalamic activity. We propose that information encoded in the high firing rate of an individual retinal ganglion cell becomes distributed among several geniculate neurons that fire synchronously. Because synchronous geniculate action potentials are highly effective in driving cortical neurons³, it is likely that information encoded by this strategy is transmitted to the next level of processing.

We made simultaneous recordings from 205 pairs of neurons in the cat; one neuron of each pair was in the retina and the other was in the lateral geniculate nucleus (LGN). Because retinal ganglion cells have high firing rates and are highly effective in driving geniculate neurons^{4–7}, this pathway is well suited for the detailed analysis of extracellular data. We mapped receptive fields using a random, computer-generated stimulus^{8–10}. An example of the receptive fields of two simultaneously recorded neurons (off-centre X cells), one in the retina and the other in the LGN, is shown in Fig. 1a, b.

The raster plot in Fig. 1c shows the occurrence of geniculate spikes relative to 300 successive ganglion-cell spikes, each centred at time zero. Following a retinal spike, there was a clear tendency for the geniculate cell to generate a spike with a peak delay of 4.8 ms—a delay consistent with a monosynaptic connection^{4,5}. This short latency peak is best appreciated in the cross-correlogram (Fig. 1d)—the average of the raster plots shown in Fig. 1c over all retinal spikes^{4,5}. Of the 205 pairs of cells in this study, 12 had statistically significant¹¹ narrow correlogram peaks, indicative of monosynaptic connections.

We next determined whether the efficacy of a retinal spike (the probability of evoking a geniculate spike) was influenced by preceding patterns of retinal activity. Patterns of firing were characterized by two temporal parameters—interspike interval and dead time—that defined a pair of retinal spikes. In each pair, the second spike followed the first by a specific interspike interval (ISI) and the first spike was preceded by a minimum period of silence (dead time). The dead time ensured a comparable level of activity immediately preceding all first spikes. For example (Fig. 1e, f), we selected all pairs of retinal spikes that occurred 10.0 ± 0.4 ms apart and followed a dead time of at least 20.0 ms. As seen either as a raster plot (Fig. 1e) or as a paired-spike cross-correlogram (Fig. 1f), second retinal spikes were significantly more effective. We call this effect paired-spike enhancement, to distinguish it from other previously described effects such as paired-pulse facilitation.

The degree of paired-spike enhancement depended on both

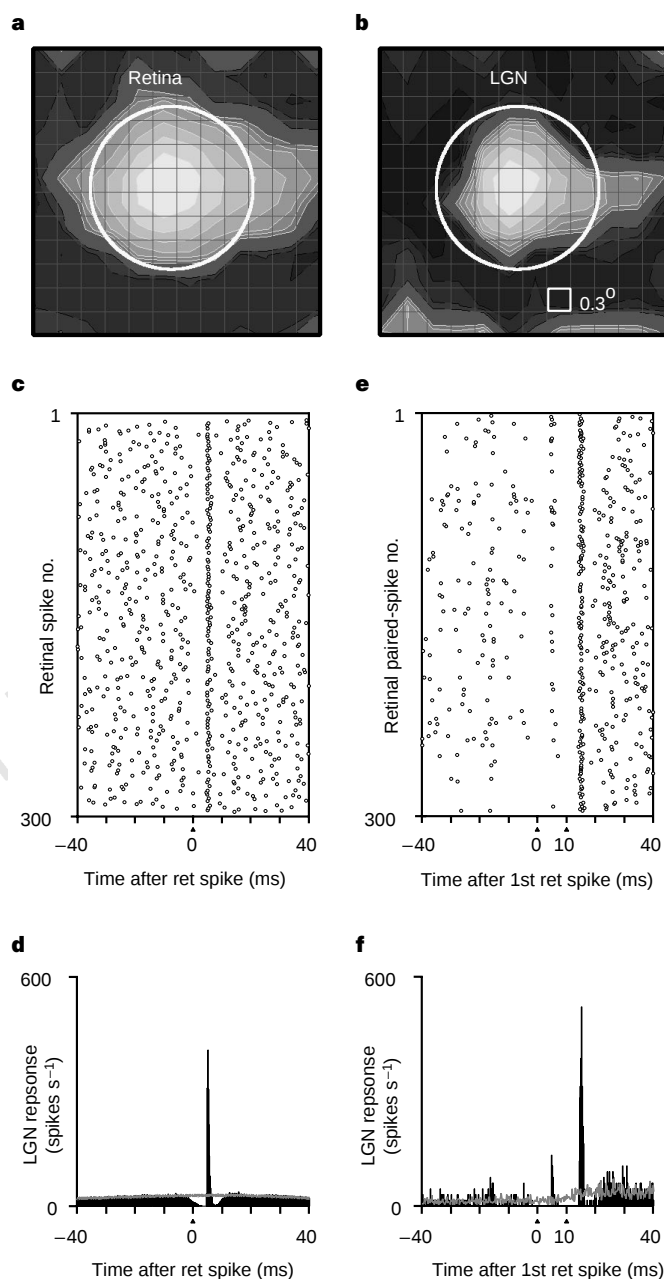


Figure 1 Recording of connected neurons in retina and LGN. **a, b**, Receptive fields of a retinal ganglion cell and geniculate neuron (on-centre X cells) mapped with white noise^{8–10}. Bright regions were excited by bright stimuli, and dark regions by dark stimuli (pixel size 0.3°). Stimulus-response delay shown: 15–31 ms for retina, 31–46 ms for LGN. Circles represent a gaussian fit to the retinal receptive field (radius: $2.5\sigma_{\text{ret}}$). **c**, Raster plot of geniculate firing (stimulated by drifting grating) relative to 300 retinal spikes. The geniculate neuron often fired ~ 5 ms after the retinal spike. **d**, Cross-correlogram, which is equivalent to the sum of all rows of the raster plot (**c**); units are LGN spikes per second following an average retinal spike. The narrow, short-latency peak (above the stimulus-dependent shuffle correlogram³⁰ (grey line) indicates that the retinal ganglion cells provided monosynaptic input to the geniculate neuron^{4,5}. Total time 797 s. Retinal spikes: 31,315; LGN spikes: 10,749. Retinal efficacy: 29.7% (see Methods). **e**, Raster plot of geniculate firing relative to 300 pairs of retinal spikes, each separated by 10.0 ± 0.2 ms and preceded by at least 20.0 ms of dead time. More geniculate spikes followed second retinal spikes than followed first retinal spikes. **f**, Cross-correlogram between paired retinal spikes and LGN. Number of retinal paired spikes is 336. Efficacy of first spikes (peak 1), 8.6%. Efficacy of second spikes (peak 2), 47.0%.

temporal parameters: interspike interval and dead time. The effect of varying the interspike interval was clear: the efficacy of second spikes reached a maximum at short intervals (~ 2 – 5 ms) and declined smoothly for intervals up to ~ 30 ms. This was found during excitation by a grating stimulus (Fig. 2a, d), by a white-noise stimulus (Fig. 2b, d), or in the absence of a visual stimulus (Fig. 2c).

We next studied the consequences of varying the dead time preceding the first spike. As would be expected, the average efficacy of first spikes (Fig. 2e) depended on the dead time preceding it; 5.4% efficacy was found when dead times were greater than 5 ms, and 1.8% efficacy was found for dead times of more than 30 ms. More important, varying the dead time also affected the peak efficacy of second spikes (Fig. 2f); efficacy decreased from 15.2% for dead times of more than 5 ms to 11.5% for dead times of more than 30 ms. Thus, the efficacy of a retinal spike is influenced not only by the immediately preceding spike, but also by the earlier level of activity. Regardless of the preceding dead time, however, the efficacy

of second spikes was always much greater than the efficacy of first spikes.

All monosynaptic connections between the retina and the LGN, both strong and weak, showed paired-spike enhancement. The scatter plot in Fig. 3 shows the efficacy of second versus first retinal spikes for each cell pair ($n = 12$), averaged over all interspike intervals between 4 and 30 ms. All but one of the points fall above the line of unit slope, indicating that second spikes were more effective than first spikes in driving a geniculate response.

Our results show that when two retinal spikes occur within ~ 3 – 30 ms of each other, the second spike has an increased probability of evoking a postsynaptic spike. Whether this effect is the result of presynaptic mechanisms, such as calcium accumulation^{2,12,13}, or postsynaptic mechanisms, such as passive integration¹⁴, cannot be determined from extracellular data. As most central synapses studied *in vitro* exhibit paired-pulse facilitation or depression (which rely on presynaptic mechanisms)^{1,2}, we predict that the effect is at least partially presynaptic. Paired-spike enhancement is unlikely to be a polysynaptic phenomenon, either cortical or intrathalamic, because of its fast onset; the effect is usually maximal at the shortest intervals that we can study, about 3 ms (determined by retinal relative refractory period). Furthermore, the effect is present without visual stimulation (Fig. 2c), when cortical feedback neurons are relatively silent. Finally, because some LGN neurons receive input from up to four ganglion cells^{4,5,15}, paired-spike enhancement might rely on other retinal inputs. This is unlikely, however, because it is seen when a single retinal ganglion cell drives almost all thalamic spikes (up to 82%; Fig. 2a–c).

Although it is likely that there is a presynaptic component to

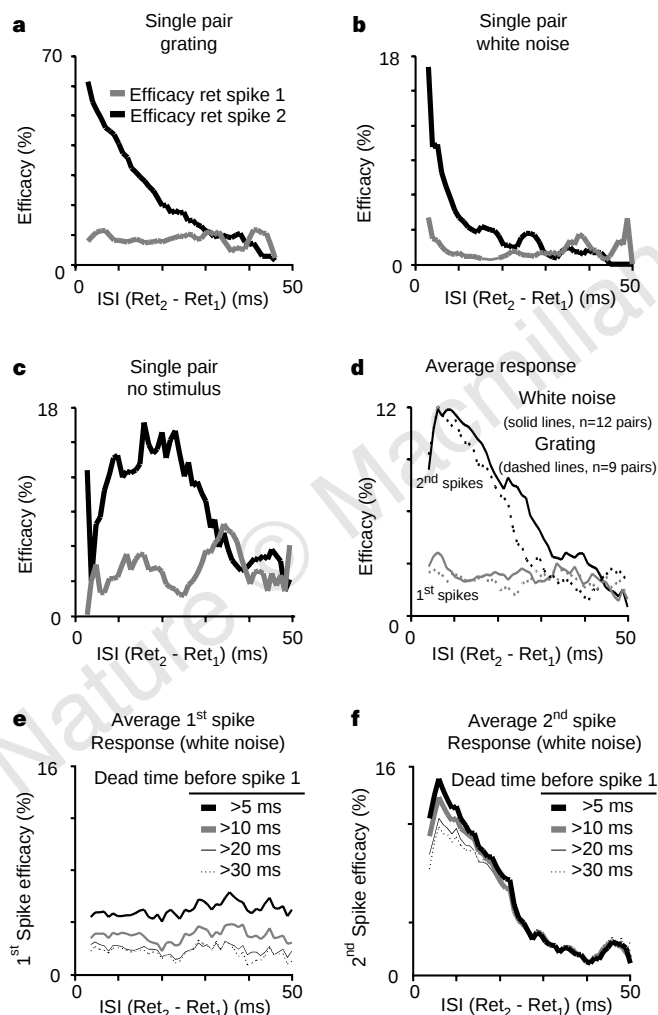


Figure 2 Time course and magnitude of paired-spike enhancement. **a–d**, Efficacy of pairs of retinal spikes (per cent that evoked a geniculate spike) that occurred at different interspike intervals (ISIs) following dead times of >20 ms. Grey lines indicate the efficacy of first retinal spikes; black lines indicate the efficacy of second retinal spikes. Efficacy at each retinal ISI was calculated as for Fig. 1, then smoothed with a 4-ms boxcar average. **a–c**, Efficacy profiles for a single retinogeniculate pair (the same pair as in Fig. 1) when **a**, excited by a grating, **b**, excited by white noise, or **c**, without visual stimulation (eyes covered). **d**, Average efficacy profiles for all pairs, excited with grating (9 pairs, dashed lines) or white noise (12 pairs, solid lines). **e, f**, Dependence of efficacies of first (**e**) and second (**f**) spikes on the dead time preceding the first spike. Curves were indistinguishable for dead times of >20 ms and >30 ms (as well as for >50 ms, not shown).

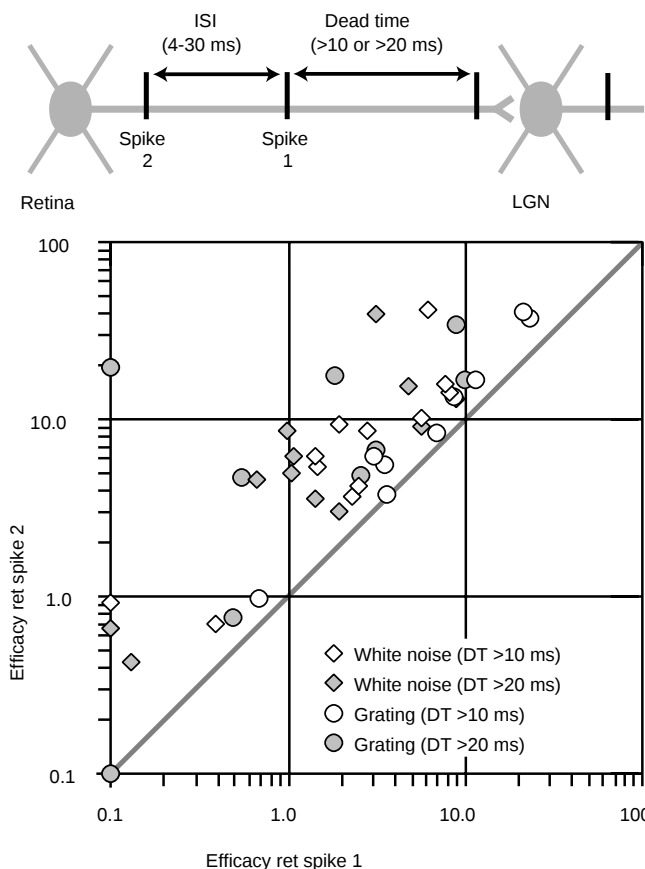


Figure 3 Scatter plot of efficacies of second versus first retinal spikes; all retinogeniculate connections showed paired-spike enhancement. Efficacies were averaged over an ISI range of 4–30 ms, weighted by number of occurrences of each ISI. Diamonds, white-noise stimuli; circles, grating stimuli. Dead time (DT) was either >10 ms (unfilled symbols) or >20 ms (filled symbols). Points on either axis are for efficacies $<0.1\%$.

paired-spike enhancement, it almost certainly involves a dynamic interplay between facilitation and depression. Most classic studies of facilitation and depression *in vitro*^{1,2} have relied on protocols in which pairs of stimuli are delivered on a quiescent baseline, or in which trains of spikes are elicited at constant rates. Several later studies^{16–19} have used more naturalistic presynaptic spike trains, similar to those found *in vivo*, and found that synaptic efficacy depends, in a complex way, on the local temporal structure of the input. Although naturalistic spike trains, delivered *in vitro*, are important in understanding the mechanisms of synaptic modulation, the functional significance of these mechanisms can only be fully appreciated *in vivo*, when they occur in the context of the normal activity of the intact nervous system.

We next determined how paired-spike enhancement affects the activity of ensembles of thalamic neurons, particularly synchronous

firing. Because retinal ganglion cells diverge anatomically to contact several thalamic targets^{15,20}, tightly synchronous activity of geniculate neurons by common input was predicted²¹ well before it was observed³. Synchronous geniculate spikes are of particular interest, as they are more effective at driving cortical responses³. Here we have found that synchronous firing in the LGN is caused by retinal divergence, but also that paired-spike enhancement increases the number of synchronous events.

Because we used a multielectrode array to record from the LGN, we were able twice to record simultaneously from two synchronized geniculate neurons and from a common retinal input. In one such example (Fig. 4), the receptive fields of one retinal cell and two simultaneously recorded geniculate cells all overlapped. The strong and narrow peaks (<1 ms) in the retinogeniculate correlograms (Ret → LGN A; Ret → LGN B) indicate that the two geniculate cells both received input from the retinal cell. As a result of this common input, the two geniculate cells fired many of their spikes synchronously (Fig. 4, LGN A–LGN B). In this case, the retinal cell drove 84% of the simultaneous spikes between LGN cells A and B.

How does paired-spike enhancement influence the relative incidence of synchronous LGN spikes? In this case, second retinal spikes were 12.3 times more likely than first spikes to evoke synchronous geniculate responses (Fig. 4, bottom). For the other three-cell recording, second retinal spikes were 1.75 times more effective in evoking synchronous geniculate spikes. Thus, as would be expected from the paired-spike effect on the activity of a single geniculate neuron, paired-spike enhancement increases the occurrence of synchronous spikes in geniculate neurons.

Correlated spiking of LGN cells is important for cortical development^{22–24}, but may also play a role in information coding (compare refs 25 and 26). In particular, synchronous geniculate spikes convey information beyond that conveyed by non-synchronous spikes²⁵. Similarly, extra information is conveyed by the retina during periods of very high spike rate²⁷. With paired-spike enhancement, this information may be selectively transmitted to the LGN, in particular to synchronous spikes. Thus there is a partial transition from a rate code to a synchronous population code.

Paired-spike enhancement, together with divergent connections from one retinal cell onto several geniculate neurons, works to increase the number of synchronous spikes in the LGN. Further, synchronous geniculate spikes can interact synergistically in evoking cortical spikes (compare ref. 3 with refs 28, 29). Thus, two aspects of the anatomy of the visual pathway—divergent input to the LGN and convergent input to the visual cortex—have physiological counterparts, namely thalamic synchrony and thalamocortical synergy, respectively. This interplay of anatomy and physiology acts not only to reinforce the pathway from periphery to cortex, but also to provide the cortex with more information about the visual environment. □

Methods

Cats were prepared for electrophysiological recordings as described^{3,11}. Recordings in the Alaminae of the dorsal LGN were made with a seven-electrode array (Thomas Recording, Marburg). Retinal recordings were made with single electrodes (AM Systems) inserted into the eye through a guide tube. Spike isolation was confirmed with off-line waveform analysis (DataWave Systems); the presence of a refractory period in autocorrelations; and observations of analogue data recorded on tape. The significance of correlogram peaks was assessed using the method of ref. 11, but with a bandpass filter of 350–2,100 Hz to capture the very fast peaks. The peak integral was determined from unfiltered correlograms (2.0-ms range around maximum), minus the baseline (taken from 2-ms ranges immediately before and after the peak range). The efficacy of retinal spikes was defined as: (peak integral)/(total retinal spikes).

Received 12 June; accepted 3 August 1998.

1. Magleby, K. L. in *Synaptic Function* (eds Edelman, G. M., Gall, W. E. & Cowan, W. M.) 21–56 (Wiley, New York, 1987).
2. Zucker, R. S. Short-term synaptic plasticity. *Annu. Rev. Neurosci.* **12**, 13–31 (1989).

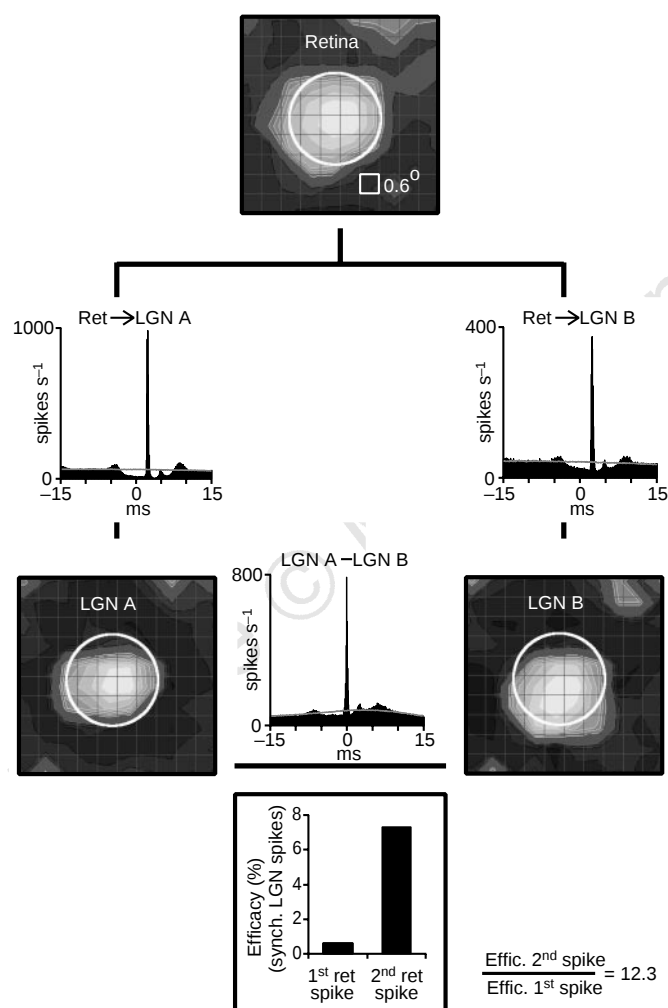


Figure 4 Divergent input from a single retinal ganglion cell synchronizes two geniculate neurons; paired-spike enhancement increases the frequency of this synchronization. The receptive fields of three simultaneously recorded on-centre Y cells (Retina, LGN A and LGN B) are shown. Circles represent the Gaussian fit to the retinal receptive field (radius: $2.5\sigma_{\text{ret}}$). Correlograms at the sides of the figure (Ret → LGN A; Ret → LGN B) have strong and narrow monosynaptic peaks, displaced ~2.5 ms to the right of zero. The lower correlogram (LGN A → LGN B) indicates that the geniculate cells fired many of their spikes synchronously (within 1.0 ms, centred at time zero). 84% of the synchronous spikes in the two geniculate cells occurred 2.5 ± 1.0 ms after a spike in the retinal cell. The histogram at the bottom shows the efficacy of first and second retinal spikes (ISI, 3–30 ms; dead time > 20 ms) in evoking synchronous geniculate spikes. Second retinal spikes were 12.3 times more likely than first spikes to evoke synchronous geniculate spikes.

3. Alonso, J.-M., Usrey, W. M. & Reid, R. C. Precisely correlated firing in cells of the lateral geniculate nucleus. *Nature* **383**, 815–819 (1996).
4. Cleland, B. G., Dubin, M. W. & Levick, W. R. Sustained and transient neurones in the cat's retina and lateral geniculate nucleus. *J. Physiol.* **217**, 473–496 (1971).
5. Cleland, B. G., Dubin, M. W. & Levick, W. R. Simultaneous recording of input and output of lateral geniculate neurones. *Nature New Biol.* **231**, 191–192 (1971).
6. Hubel, D. H. & Wiesel, T. N. Integrative action in the cat's lateral geniculate body. *J. Physiol.* **155**, 385–398 (1961).
7. Kaplan, E., Purpura, K. & Shapley, R. M. Contrast affects the transmission of visual information through the mammalian lateral geniculate nucleus. *J. Physiol.* **391**, 267–288 (1987).
8. Reid, R. C., Victor, J. D. & Shapley, R. M. The use of m-sequences in the analysis of visual neurons: linear receptive field properties. *Vis. Neurosci.* **14**, 1015–1027 (1997).
9. Sutter, E. E. in *Advanced Methods of Physiological Systems Modeling* Vol. 1 (ed. Marmarelis, V.) 303–315 (Univ. Southern California, Los Angeles, 1987).
10. Sutter, E. E. in *Nonlinear Vision: Determination of Neural Receptive Fields, Function and Networks* (eds Pinter, R. & Nabet, B.) 171–220 (CRC, Cleveland, 1992).
11. Reid, R. C. & Alonso, J.-M. Specificity of monosynaptic connections from thalamus to visual cortex. *Nature* **378**, 281–284 (1995).
12. Katz, B. & Miledi, R. The role of calcium in neuromuscular facilitation. *J. Physiol.* **195**, 481–492 (1968).
13. Atluri, P. P. & Regehr, W. G. Determinants of the time course of facilitation at the granule cell to Purkinje cell synapse. *J. Neurosci.* **16**, 5661–5671 (1996).
14. Bloomfield, S. A., Hmaos, J. E. & Sherman, S. M. Passive cable properties and morphological correlates of neurones in the lateral geniculate nucleus of the cat. *J. Physiol.* **383**, 653–692 (1987).
15. Hamos, J. E., Van Horn, S. C., Raczkowski, D. & Sherman, S. M. Synaptic circuits involving an individual retinogeniculate axon in the cat. *J. Comp. Neurol.* **259**, 165–192 (1987).
16. Varela, J. A. *et al.* A quantitative description of short-term plasticity at excitatory synapses in layer 2/3 of rat primary visual cortex. *J. Neurosci.* **17**, 7926–7940 (1997).
17. Tsodyks, M. V. & Markram, H. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proc. Natl Acad. Sci. USA* **94**, 719–723 (1997).
18. Markram, H. & Tsodyks, M. Redistribution of synaptic efficacy between neocortical pyramidal neurons. *Nature* **382**, 807–810 (1996).
19. Abbott, L. F., Varela, J. A., Sen, K. & Nelson, S. B. Synaptic depression and cortical gain control. *Science* **275**, 220–224 (1997).
20. Peters, A. & Payne, B. R. Numerical relationships between geniculocortical afferents and pyramidal cell modules in cat primary visual cortex. *Cereb. Cortex* **3**, 69–78 (1993).
21. Cleland, B. G. in *Visual Neuroscience* (eds Pettigrew, J. D., Snaderson, K. S. & Levick, W. R.) 111–120 (Cambridge Univ. Press, London, 1986).
22. Meister, M., Wong, R. O., Baylor, D. A. & Shatz, C. J. Synchronous bursts of action potentials in ganglion cells of the developing mammalian retina. *Science* **252**, 939–943 (1991).
23. Miller, K. D. A model for the development of simple cell receptive fields and the ordered arrangement of orientation columns through activity-dependent competition between ON- and OFF-center inputs. *J. Neurosci.* **14**, 409–441 (1994).
24. Weliky, M. & Katz, L. C. Disruption of orientation tuning in visual cortex by artificially correlated neuronal activity. *Nature* **386**, 680–688 (1997).
25. Dan, Y., Alonso, J.-M., Usrey, W. M. & Reid, R. C. Coding of visual information by precisely correlated spikes in the lateral geniculate nucleus. *Nature Neurosci.* (in the press).
26. Meister, M., Lagnado, L. & Baylor, D. A. Concerted signaling by retinal ganglion cells. *Science* **270**, 1207–1210 (1995).
27. Smirnakis, S. M., Warland, D. K., Berry, M. J. & Meister, M. Spike bursts in visual responses of retinal ganglion cells. *Soc. Neurosci. Abstr.* **22**, 494 (1996).
28. König, P., Engel, A. K. & Singer, W. Integrator or coincidence detector? The role of the cortical neuron revisited. *Trends Neurosci.* **19**, 130–137 (1996).
29. Usrey, W. M. & Reid, R. C. Synchronous activity in the visual system. *Annu. Rev. Physiol.* (in the press).
30. Perkel, D. H., Gerstein, G. L. & Moore, G. P. Neuronal spike trains and stochastic point processes. II. Simultaneous spike trains. *Biophys. J.* **7**, 419–440 (1967).

Acknowledgements. This work was supported by the NIH, The Klingenstein Fund, The Harvard Mahoney Neuroscience Institute, and The Howard Hughes Medical Institute. We thank E. Serra for technical assistance and J.-M. Alonso, J. Assad, M. Livingstone, M. Meister, W. Regehr and D. Hubel for comments on this manuscript.

Correspondence and requests for materials should be addressed to R.C.R. (e-mail: clay_reid@hms.harvard.edu).

***Drosophila* oocyte localization is mediated by differential cadherin-based adhesion**

Dorothea Godt & Ulrich Tepass

Department of Zoology, University of Toronto, 25 Harbord Street, Toronto, Ontario M5S 3G5, Canada

These authors contributed equally to this work

In a *Drosophila* follicle the oocyte always occupies a posterior position among a group of sixteen germline cells. Although the importance of this cell arrangement for the subsequent formation of the anterior–posterior axis of the embryo is well documented^{1–4}, the molecular mechanism responsible for the posterior localization of the oocyte was unknown. Here we show that the

homophilic adhesion molecule *DE-cadherin*^{5–7} mediates oocyte positioning. During follicle biogenesis, *DE-cadherin* is expressed in germline (including oocyte) and surrounding follicle cells, with the highest concentration of *DE-cadherin* being found at the interface between oocyte and posterior follicle cells. Mosaic analysis shows that *DE-cadherin* is required in both germline and follicle cells for correct oocyte localization, indicating that germline–soma interactions may be involved in this process. By analysing the behaviour of the oocyte in follicles with a chimaeric follicular epithelium, we find that the position of the oocyte is determined by the position of *DE-cadherin*-expressing follicle cells, to which the oocyte attaches itself selectively. Among the *DE-cadherin* positive follicle cells, the oocyte preferentially contacts those cells that express higher levels of *DE-cadherin*. On the basis of these data, we propose that in wild-type follicles the oocyte competes successfully with its sister germline cells for contact to the posterior follicle cells, a sorting process driven by different concentrations of *DE-cadherin*. This is, to our knowledge, the first *in vivo* example of a cell-sorting process that depends on differential adhesion mediated by a cadherin.

DE-cadherin, encoded by the gene *shotgun* (*shg*), is the major epithelial cadherin in *Drosophila* and forms a cell adhesion complex with Armadillo/ β -catenin and $\Delta\alpha$ -catenin^{5–9}. During oogenesis, *DE-cadherin* is expressed in all germline and follicle cells in the germarium and the follicles (Fig. 1c). Before follicle formation, the oocyte accumulates more *DE-cadherin* than the other 15 germline cells (the nurse cells), which are connected to the oocyte by cytoplasmic bridges known as the ring canals. The higher level of *DE-cadherin* in the oocyte depends on oocyte specification as, in *egalitarian* (*egl*) mutants, in which all 16 germline cells develop into nurse cells^{10,11}, there are no differences in *DE-cadherin* concentrations between germline cells (Fig. 1f).

Follicle formation is initiated when follicle cells grow inwards at the posterior side of the germ cell cluster, which forms a monolayer resting on the follicle cells (Fig. 1a, b). Subsequently, the germ cell cluster rounds off and eventually becomes covered with follicle cells at its anterior side also. The oocyte, which initially has a variable position, always assumes the most posterior position during follicle formation. Higher levels of *DE-cadherin* are seen in the posterior and anterior follicle cells than in lateral follicle cells (Fig. 1d). However, the highest *DE-cadherin* concentration is seen at the interface between the oocyte and the posterior follicle cells (Fig. 1d, e). In *egl* mutants, anterior and posterior follicle cells also express increased *DE-cadherin* levels, but no anterior–posterior differences are seen (Fig. 1f). Increased *DE-cadherin* concentrations at the interface between the oocyte and posterior follicle cells are maintained until stage 7 of oogenesis. Armadillo is distributed similarly to *DE-cadherin* (Fig. 1g). The *DE-cadherin* expression pattern, together with the finding that *shg* mutant germ line clones fail to produce eggs⁶, is consistent with a function for this adhesion molecule in both germline and follicle cells. In particular, the high accumulation of *DE-cadherin* at the interface of oocyte and posterior follicle cells indicates that it may mediate interactions between these two cell types.

To study the consequences of loss of *DE-cadherin* expression during oogenesis, we first analysed the semiviable allelic combination *shg*^{P34-1}/*shg*^{R6}. Mislocalization of the oocyte was seen in 29.5% of stage 7–9 mutant follicles ($n = 383$), compared with 0% in wild-type follicles ($n = 498$) (Fig. 2). To determine whether *DE-cadherin* is required in the germline or in the follicular epithelium for oocyte localization, and to analyse oogenesis in the absence of *DE-cadherin* expression, we generated cell clones homozygous for the null allele *shg*^{R69}. The most prominent phenotype of germline clones that lack *DE-cadherin* is a mislocalization of the oocyte (Fig. 3a, b, e, f). An abnormal position of the oocyte is seen already in stage 1 follicles and in 89% of follicles with a mutant germline (Fig. 3i). In addition to mislocalization of the oocyte, lack of *DE-cadherin* expression

leads to several other defects in the germline during oogenesis, such as follicles with more germ cells, abnormally shaped germ cells, fused nurse cells, ectopic cytoplasmic accumulations of F-actin, and incomplete transport of nurse cell contents into the oocyte (data not shown). These phenotypic traits are variable in expressivity and occur with incomplete penetrance. Similar defects have been described for germline clones of *armadillo* (*arm*)¹² and *shg* (ref. 13) mutations. Another study of *shg* germline clones did not reveal defects in oocyte positioning¹⁴. In contrast to these studies, which described late stages of oogenesis, we find that *DE*-cadherin is already required for oocyte positioning during follicle formation, at the time at which the oocyte establishes its posterior position

(Fig. 3b). The similarity of the defects seen in *arm* and *shg* germline clones indicates that these defects may result from lack of function of the *DE*-cadherin–catenin complex.

Oocyte mislocalization from stage 1 onwards is also seen in most follicles that express *DE*-cadherin in the germline but do not express it in the follicular epithelium (68%; Fig. 3c, g, i). Finally, mislocalization of the oocyte is also observed in follicles that are mutant for *shg* in both germline and soma (Fig. 3d, h). All *shg* mutant follicles eventually degenerate at different stages during oogenesis. In all of our experiments the oocyte was specified, indicating that *DE*-cadherin does not function in oocyte specification but is directly involved in oocyte positioning. Direct cell–cell contact between the

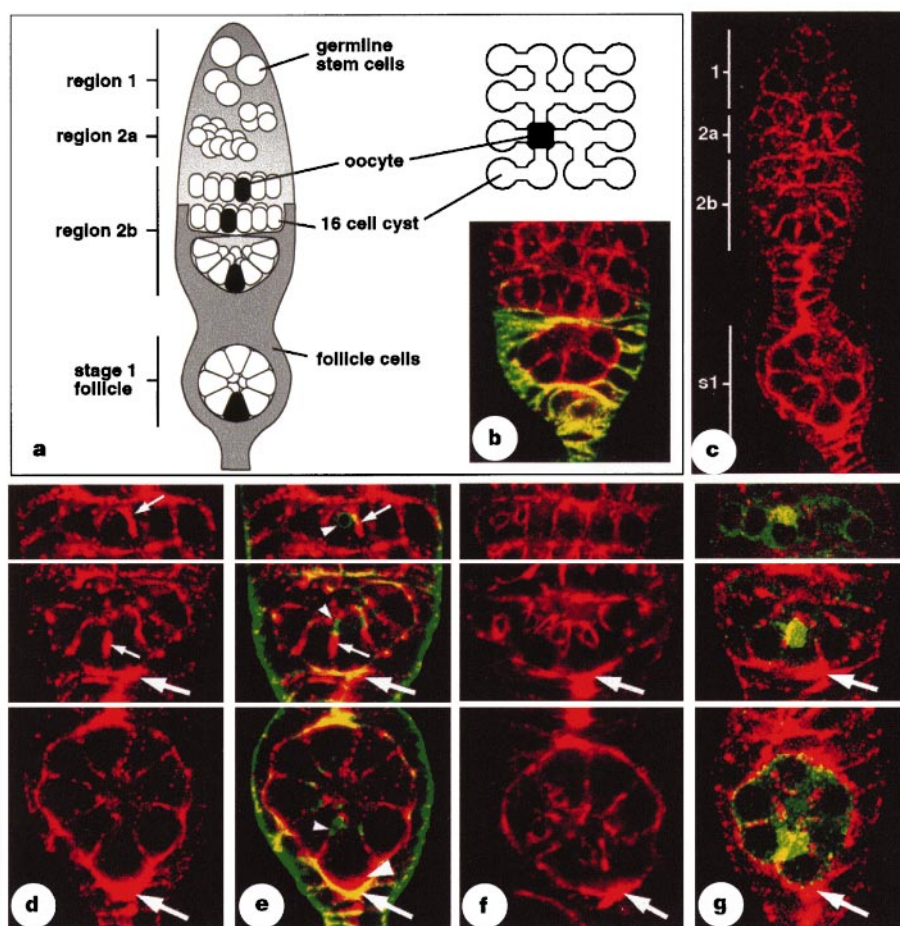


Figure 1 Distribution of *DE*-cadherin and Armadillo in the germarial region of the ovary. **a**, A germarium. Offspring of the germline stem cells divide four times and give rise to cysts of 16 cells each, connected by ring canals³⁰. Two cyst cells, one of which develops into the oocyte, have four ring canals each. The cyst forms a monolayer in region 2b with the oocyte located in the central area, and continues to develop, through a cup-shaped intermediate, into a round cluster, the stage 1 follicle. During this process the cyst is ensheathed by follicle cells which rearrange into an epithelium. The oocyte has established its posterior position at stage 1. **b**, In region 2b, follicle cells that express Fasciclin III (green; *DE*-cadherin expression is in red) penetrate between 16-cell cysts and first contact the posterior side of the cyst. **c**, *DE*-cadherin (red) is expressed in both germline and follicle cells; s1, stage 1 follicle. **d–g**, Three phases in the formation of a follicle: monolayered cyst (upper panels), cup-shaped intermediate (middle panels), and stage 1 follicle (lower panels). **d, e**, In region 2b, *DE*-cadherin (red) concentrates at the cell membranes between cyst cells adjacent to the ring canals (upper and middle panels). Ring canals are highlighted by F-actin staining (green). Particu-

larly intense staining is seen between the two four-ring canal cells (small arrows) that share the largest ring canal (small arrowheads in **e**). High concentrations of *DE*-cadherin are found in posterior follicle cells (large arrows) and opposing anterior follicle cells. In stage 1 follicles, the concentration of *DE*-cadherin is reduced at the membranes between cyst cells. A prominent crescent of *DE*-cadherin forms in the oocyte at the surface that contacts the posterior follicle cells in wild-type follicles (large arrowhead in **e**), whereas lower concentrations of *DE*-cadherin are seen at the interface between nurse cells and follicle cells. **f**, In *egl^{4e}/Df(2R)bw^{S46}* mutants no asymmetric distribution of *DE*-cadherin is seen between germline cells. Posterior (arrows) and opposing anterior follicle cells accumulate *DE*-cadherin as in the wild type, but there are no anterior–posterior differences. **g**, The distribution of Armadillo (red) is similar to that of *DE*-cadherin. Arrows point to the accumulation of Armadillo at the interface of oocyte and posterior follicle cells. Egalitarian (green) is used to identify the oocyte; it is expressed in all germline cells but accumulates in the oocyte¹¹. Anterior is shown at the top.

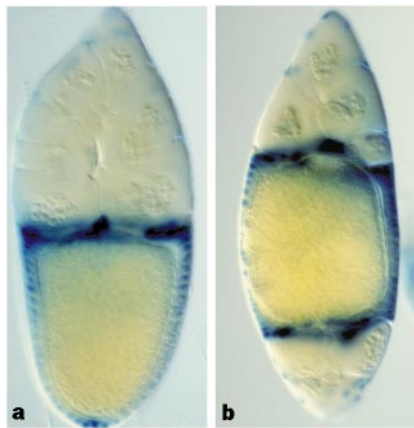


Figure 2 Oocyte mislocalization in *shg* mutant follicles. **a**, Wild-type stage 10 follicle with posteriorly localized oocyte. **b**, *shg^{R6}/shg^{P34-1}* stage 10 follicle with centrally located oocyte. β -galactosidase expression of the *P-lacW* insertion of *shg^{P34-1}* was assayed by X-gal staining. This reporter shows a *shg*-specific distribution in follicle cells but is not expressed in germline cells. Anterior follicle cells express particularly high levels. When the oocyte is located centrally both the anterior and the posterior follicle cells adopt an anterior fate². Anterior is shown at the top.

oocyte and the posterior follicle cells is essential for formation of the anterior–posterior axis². Centrally located oocytes in *shg* mutant follicles consistently fail to properly localize anterior and posterior determinants, whereas most of the anteriorly located oocytes show a normal distribution of determinants in reversed orientation (Fig. 4). Thus our results indicate that *DE*-cadherin-dependent homophilic adhesion between the follicular epithelium and the oocyte is required for oocyte positioning.

The position of the oocyte within follicles that contain either *shg* germline or *shg* follicle cell clones is not completely random but shows a bias towards posterior (Fig. 3i). This process appears to be independent of oocyte specification, as preferential posterior localization of the two cells with four ring canals (one of which normally becomes the oocyte) is also seen in *egl* (ref. 10) or *Bicaudal-D* (ref. 15) mutants, in which oocyte specification is blocked. The posterior bias of the oocyte might result from constraints imposed by the spatial arrangement of the cyst cells during follicle formation when the centrally located cells that include the oocyte are pushed towards posterior (Fig. 1a, g)¹⁶. However, as the posterior bias is also observed in the absence of oocyte specification it can be ruled out as the main mechanism that generates the posterior positioning of the oocyte, because such a mechanism needs to distinguish between oocyte and nurse cells.

We have shown that *DE*-cadherin is required in both germline

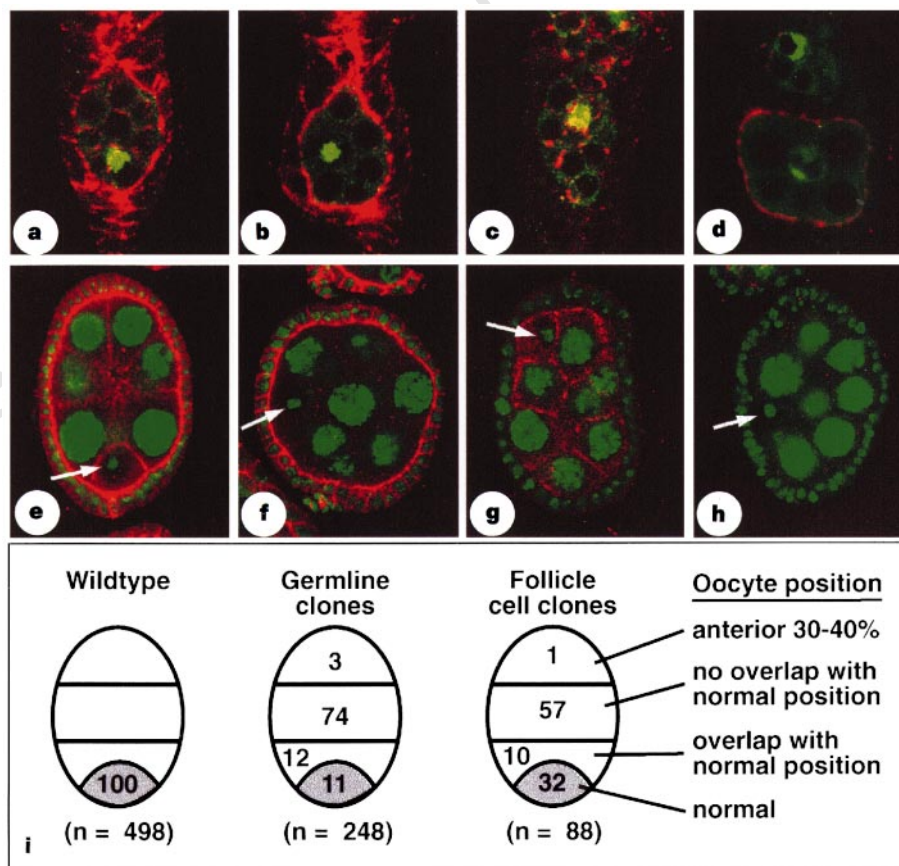


Figure 3 Oocyte localization in follicles with a *shg^{R69}* mutant germline or follicular epithelium. **a–d**, Stage 1 follicles. Lack of Armadillo (Armadillo is stained in red) marks *shg* mutant clones. Egalitarian staining (green) reveals the oocyte position. **a**, Wild-type follicle with a posterior oocyte. **b**, Germline mutant clone and **c**, follicle cell mutant clone with mislocalized oocytes. **d**, Oocyte mislocalization in stage 1 (upper) and stage 4 (lower) follicles with mutant germline and follicular epithelium. *shg* mutant germline clones do not accumulate detectable levels of Armadillo, whereas *shg* mutant follicular epithelia retain a small amount of Armadillo at the zonula adherens, indicating that a cadherin–catenin complex that does not

involve *DE*-cadherin may be present. **e–h**, Stage 6 or 7 follicles. Lack of *DE*-cadherin (*DE*-cadherin is stained in red) identifies mutant clones. Use of Picogreen (green) allows to identify the small oocyte nucleus (arrows). **e**, Wild-type. There is *DE*-cadherin expression in both germline and follicle cells. **f**, Germline mutant clone. **g**, Follicle cell mutant clone. **h**, Follicle with mutant germline and follicular epithelium. **i**, Summary of oocyte position in mutant germline and follicle cell clones. The numbers show the percentage of follicles with an oocyte located at the indicated position. Anterior is shown at the top.

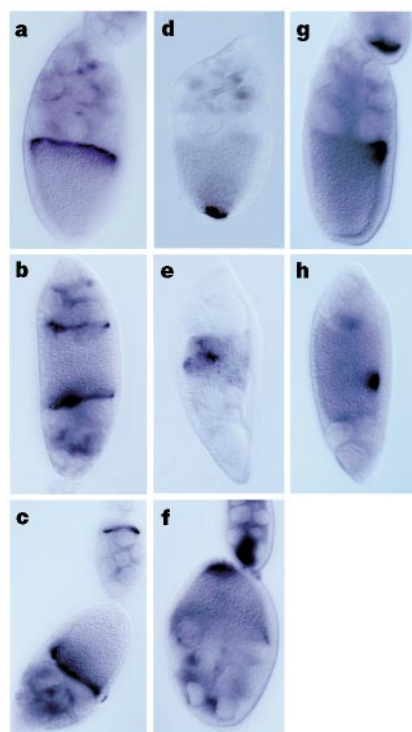


Figure 4 Oocyte polarity in *shg* mutant follicles. Messenger RNA distribution of anterior (*bicoid*, **a–c**), posterior (*oskar*, **d–f**), and dorsal (*gurken*, **g, h**) determinants in wild-type follicles (**a, d, g**), and in *shg* mutant follicles with a central oocyte (**b, e, h**) or an anterior oocyte (**c, f**). We did not determine whether oocyte mislocalization was caused by the lack of *shg* expression in the germline or in the follicle cells. All three determinants are mislocalized in follicles with a centrally located oocyte. Follicles with an anterior oocyte show a reversed but otherwise normal distribution of *bicoid* and *oskar*, indicating that the anterior follicle cells have the same potential to interact with the oocyte as do the posterior follicle cells. Anterior is shown at the top.

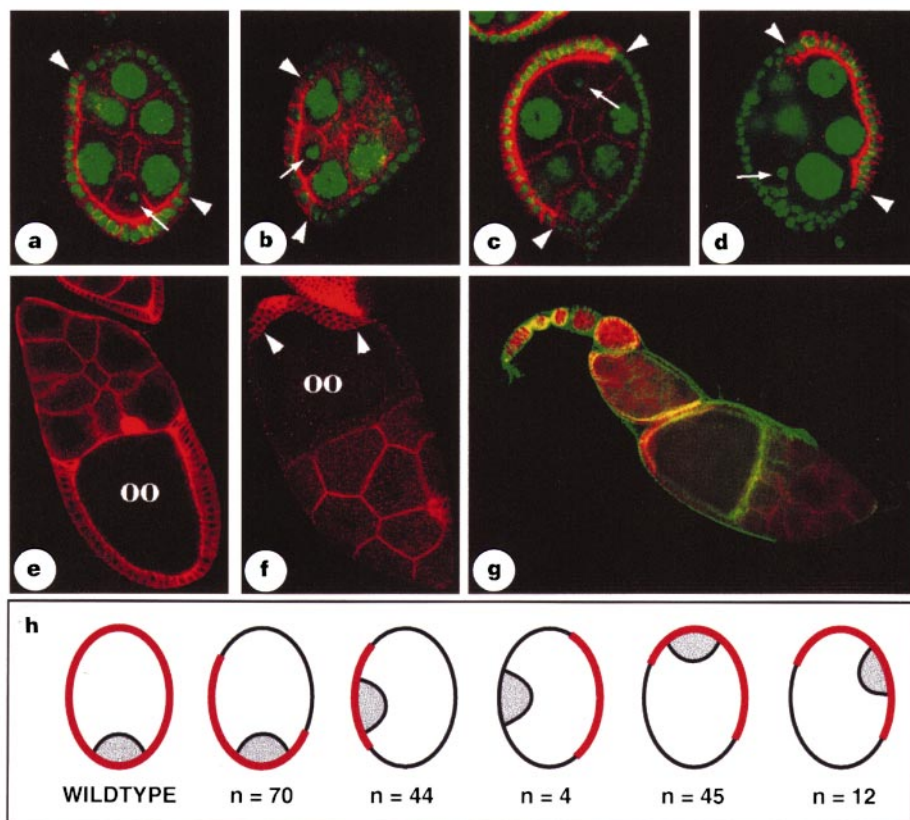


Figure 5 Oocyte localization in follicles with an *shg* chimaeric follicular epithelium. **a–d**, Absence of *DE-cadherin* (*DE-cadherin* is stained in red) reveals the extent of the *shg*^{R69} mutant follicle cell clone (arrowheads). Picogreen (green) allows us to identify the small oocyte nucleus (arrows). **a–c**, The oocyte contacts *DE-cadherin*-expressing follicle cells. **c**, Follicle with anterior and lateral *DE-cadherin*-positive follicle cells in which the oocyte occupies the most anterior position. **d**, Follicle with an *shg* chimaeric follicular epithelium and an *shg* mutant germline. The oocyte is not attached to *DE-cadherin*-expressing follicle cells. **e**, Wild-type stage 10 follicle, showing the posteriorly located oocyte (oo). *DE-cadherin* (red) is expressed in all follicle cells. **f**, Mosaic stage 10 follicle showing an anterior patch of *DE-cadherin* (red) positive follicle cells (arrowheads). Note the anteriorly located oocyte. **g**, Ovariole containing the follicle seen in (**f**) to show its orientation. F-actin staining (green) highlights the myoepithelial sheath of the ovariole. **h**, Summary of oocyte distribution in follicles with an *shg* chimaeric follicular epithelium. Only follicles that contain *DE-cadherin*-positive germline cells are shown. The red line indicates the approximate extent of the *DE-cadherin*-expressing follicle cell patch. Anterior is shown at the top in all panels, except in **g** in which anterior is to the left.

and follicle cells to establish the posterior position of the oocyte. Posterior oocyte localization cannot, however, be explained by a simple homophilic cadherin-mediated recognition process between oocyte and posterior follicle cells, because *DE-cadherin* is not restricted to those cells but is expressed ubiquitously in both germline and follicle cells. Initially, the oocyte has a variable position among the central cells of the monolayered 16-cell cyst, but it achieves an invariable posterior position during the transition to the stage 1 follicle. During this time, oocyte and posterior follicle cells have higher levels of *DE-cadherin* than surrounding cells (Fig. 1d, e), and the distribution of *DE-cadherin* in *egl* mutants indicates that the accumulation of *DE-cadherin* in both cell types occurs independently of each other (Fig. 1f). This raises the possibility that the movement of the oocyte to the posterior pole of a follicle is driven by a sorting mechanism that relies on different concentrations of *DE-cadherin*. To test this hypothesis experimentally, we studied oocyte localization in chimaeric follicles in which 50% or fewer of the follicle cells retained *DE-cadherin* expression. We found that the oocyte attaches itself with high fidelity (98%; $n = 175$) to the remaining *DE-cadherin*-expressing follicle cells, independently of where these follicle cells are located (Fig. 5). In contrast, in only 39% of the follicles ($n = 33$) that contain a chimaeric follicular epithelium and a *shg* mutant germline is the oocyte found in contact with *DE-cadherin*-expressing follicle cells (Fig. 5d), implying that oocytes that lack *DE-cadherin* show no preferences for *DE-cadherin*-positive follicle cells. These results indicate that, in a germ cell cluster that expresses *DE-cadherin*, sorting occurs by which the oocyte successfully competes with nurse cells for contact to *DE-cadherin*-expressing follicle cells.

In chimaeric follicles that contained *DE-cadherin*-expressing posterior follicle cells, the oocyte was always found in the most posterior position, like in the wild-type follicle (Fig. 5a). If differential adhesion is responsible for the oocyte to target the posterior follicle cells, the oocyte should be able to distinguish not only between *DE-cadherin*-positive and -negative follicle cells, but also between follicle cells with different concentrations of *DE-cadherin*.

We therefore determined whether, in the absence of *DE*-cadherin-expressing posterior follicle cells, the oocyte would preferentially contact the anterior follicle cells, which also express higher levels of the protein than lateral follicle cells (Fig. 1d, e). In 79% of follicles ($n = 57$) that contain anterior and lateral but not posterior *DE*-cadherin-expressing follicle cells, the oocyte attached to the anterior follicle cells and occupied the most anterior position (Fig. 5c, f–h). Thus in each of the different types of chimaeric follicles a cell arrangement develops that allows for the highest level of homophilic adhesion between germline cells and follicle cells. These data indicate that the normal differences in levels of *DE*-cadherin between oocyte and nurse cells are sufficient to promote cell sorting so that the oocyte can actively move to, and attach itself selectively to the follicle cells that express higher amounts of *DE*-cadherin than other follicle cells.

Our results contribute on the one hand to our understanding of the mechanisms that generate anterior–posterior polarity in *Drosophila* oogenesis, and on the other to the mechanisms that control the sorting of cells. The ability of embryonic cells to sort out may be one of the main factors that is required for tissue formation and maintenance^{17,18}. The sorting of cells that express different adhesion molecules of the cadherin family or that rely on different expression levels of the same cadherin has been described for cell culture systems^{19–21}, but whether such processes act in normal development remained questionable until now. In contrast to classic *in vitro* sorting experiments, in which sorting is thought to be promoted by homotypic adhesion^{22,23}, sorting among the germline cyst cells is promoted by heterotypic adhesion between the germline cells and the follicle cells. Heterotypic interactions in this case are essential as they allow the sorting cell types, that is, the oocyte and the nurse cells, to assume a stereotypic spatial arrangement. Anterior oocyte positioning in chimaeric follicles that lack *DE*-cadherin in posterior follicle cells shows that *DE*-cadherin-mediated sorting can overcome the posterior bias in oocyte positioning seen in the absence of *DE*-cadherin in all follicle cells. Although in wild-type follicles both anterior and posterior follicle cells express higher levels of *DE*-cadherin than lateral follicle cells, the oocyte always occupies a posterior position, indicating that, during follicle formation, oocyte positioning occurs before the germ cell cluster makes contact with anterior follicle cells. As soon as contact to posterior follicle cells is established, strong *DE*-cadherin-mediated adhesion between oocyte and posterior follicle cells is likely to maintain the posterior oocyte position. We propose that *DE*-cadherin-based cell sorting is the main force that causes posterior oocyte localization and, therefore, initiates anterior–posterior axis formation in *Drosophila*. □

Methods

Alleles. *shg* alleles used were *shg*^{P34-1}, an allele of intermediate strength that carries a P-lacW insertion in the 5′-untranslated region of the *shg* gene⁶, and *shg*^{R6} and *shg*^{R69}, which were generated from *shg*^{P34-1} by P-element excisions. *shg*^{R6} is a weak, homozygous viable loss-of-function allele. *shg*^{R69} is a null allele according to several criteria. *shg*^{R69} causes a consistent strong embryonic phenotype in homozygosis and hemizygosis, and *shg*^{R69} mutant cells do not express detectable levels of *shg* messenger RNA or protein. Molecular analysis shows that *shg*^{R69} contains a deletion that extends from the mini-white gene in P-lacW into the *shg* coding region, removing the translation start site, the signal peptide and all cadherin repeats. The genetic background for all three alleles was exchanged in recombination experiments. Genotypes of *egl* mutants were *egl*^{4c}/*leg*^{WUSO} and *egl*^{4c}/*Df(2R)bw*^{S46} (ref. 11).

Clones. *shg*^{R69} mutant clones were generated by mitotic recombination with the FLP/FRT system²⁴. We used the FRT insertion *P[ry⁺;hs-neo;FRT]42D* and an X-chromosomal insertion of a P-element containing the FLP-recombinase gene under the control of a heat-shock promoter. FLP-recombinase expression was induced by heat shocks at 37 °C for 2 h at early and late second larval instar to produce *shg* mutant germline and follicle stem cells.

Antibody staining and tissue *in situ* hybridization. Antibodies used in this study were: anti-*DE*-cadherin (DCAD-2; 1:50) (ref. 5); anti-Armadillo (N2-7A1; 1:100) (ref. 25); anti-Egalitarian (1:1,000) (ref. 11); and anti-Fasciclin III (7G10; 1:25) (ref. 26). Secondary antibodies were conjugated with either Cy3 or fluorescein isothiocyanate (Jackson Laboratories). DNA was stained with Picogreen (1:1,000; Molecular Probes) and F-actin was labelled with Oregon Green 488-phalloidin (1:20; Molecular Probes). Tissue *in situ* hybridizations were done according to standard protocols using digoxigenin-labelled cDNAs of *bicoid*²⁷, *oskar*²⁸ and *gurken*²⁹.

Received 13 May; accepted 10 July 1998.

1. Ray, R. & Schüpach, T. Intercellular signaling and the polarization of body axes during *Drosophila* oogenesis. *Genes Dev.* **10**, 1711–1723 (1996).
2. González-Reyes, A. & St Johnston, D. Role of oocyte position in establishment of anterior–posterior polarity in *Drosophila*. *Science* **266**, 639–642 (1994).
3. Roth, S., Neumann-Silberberg, F. S., Barcelo, G. & Schüpach, T. *cornichon* and the EGF receptor signaling process are necessary for both anterior–posterior and dorsal–ventral pattern formation in *Drosophila*. *Cell* **81**, 967–977 (1995).
4. González-Reyes, A., Elliott, H. & St Johnston, D. Polarization of both major body axes in *Drosophila* by *gurken*-torpedo signalling. *Nature* **375**, 654–658 (1995).
5. Oda, H., Uemura, T., Harada, Y., Iwai, Y. & Takeichi, M. A *Drosophila* homolog of cadherin associated with *armadillo* and essential for embryonic cell–cell adhesion. *Dev. Biol.* **165**, 716–726 (1994).
6. Tepass, U. *et al.* *shotgun* encodes *Drosophila* E-cadherin and is preferentially required during cell rearrangement in the neuroectoderm and other morphogenetically active epithelia. *Genes Dev.* **10**, 672–685 (1996).
7. Uemura, T., Oda, H., Kraut, R., Hayashi, S. & Takeichi, M. Zygotic *Drosophila* E-cadherin expression is required for processes of dynamic epithelial cell rearrangement in the *Drosophila* embryo. *Genes Dev.* **10**, 659–671 (1996).
8. Oda, H., Uemura, T., Shiomi, T., Nagafuchi, A., Tsukita, S. & Takeichi, M. Identification of a *Drosophila* homologue of α -catenin and its association with the *armadillo* protein. *J. Cell Biol.* **121**, 1133–1140 (1993).
9. Peifer, M. The product of the *Drosophila* segment polarity gene *armadillo* is part of a protein complex resembling the vertebrate adherens junction. *J. Cell Sci.* **105**, 993–1000 (1993).
10. Carpenter, A. T. *egalitarian* and the choice of cell fates in *Drosophila melanogaster* oogenesis. *Ciba Found. Symp. Germline Dev.* **182**, 223–246 (1994).
11. Mach, J. M. & Lehmann, R. An Egalitarian–Bicardal–D complex is essential for oocyte specification and axis determination in *Drosophila*. *Genes Dev.* **11**, 423–435 (1997).
12. Peifer, M., Orsulic, S., Sweeton, D. & Wieschaus, E. A role for the *Drosophila* segment polarity gene *armadillo* in cell adhesion and cytoskeletal integrity during oogenesis. *Development* **118**, 1191–1207 (1993).
13. White, P., Aberle, H. & Vincent, J.-P. Signaling and adhesion activities of mammalian β -catenin and plakoglobin in *Drosophila*. *J. Cell Biol.* **140**, 183–195 (1998).
14. Oda, H., Uemura, T. & Takeichi, M. Phenotypic analysis of null mutants of *DE*-cadherin and *Armado* in *Drosophila* ovaries reveals distinct aspects of their functions in cell adhesion and cytoskeletal organization. *Genes to Cells* **2**, 29–40 (1997).
15. Ran, B., Bopp, R. & Suter, B. Null alleles reveal novel requirements for *Bic-D* during *Drosophila* oogenesis and zygotic development. *Development* **120**, 1233–1242 (1994).
16. Mahowald, A. P. & Strassheim, J. M. Intercellular migration of centrioles in the germline of *Drosophila melanogaster*. An electron microscopic study. *J. Cell Biol.* **45**, 306–320 (1970).
17. Takeichi, M. Cadherin cell adhesion receptors as a morphogenetic regulator. *Science* **251**, 1451–1455 (1991).
18. Gumbiner, B. M. Cell adhesion: the molecular basis of tissue architecture and morphogenesis. *Cell* **84**, 345–357 (1996).
19. Nose, A., Nagafuchi, A. & Takeichi, M. Expressed recombinant cadherins mediate cell sorting in model systems. *Cell* **54**, 993–1001 (1988).
20. Friedlander, D. R., Mege, R.-M., Cunningham, B. A. & Edelman, G. M. Cell sorting-out is modulated by both the specificity and amount of different cell adhesion molecules (CAMs) expressed on cell surfaces. *Proc. Natl Acad. Sci. USA* **86**, 7043–7047 (1989).
21. Steinberg, M. S. & Takeichi, M. Experimental specification of cell sorting, tissue spreading, and specific spatial patterning by quantitative differences in cadherin expression. *Proc. Natl Acad. Sci. USA* **91**, 206–209 (1994).
22. Townes, P. L. & Holtfrete, J. Directed movements and selective adhesion of embryonic amphibian cells. *J. Exp. Zool.* **128**, 53–120 (1995).
23. Steinberg, M. S. Adhesion in development: a historical overview. *Dev. Biol.* **180**, 377–388 (1996).
24. Xu, T. & Harrison, S. D. in *Drosophila melanogaster: Practical Uses in Cell and Molecular Biology* (eds Goldstein, L. S. B. & Fyrberg, E. A.) 655–682 (Academic, San Diego, 1994).
25. Peifer, M., Sweeton, D., Casey, M. & Wieschaus, E. wingless signal and Zeste-white 3 kinase trigger opposing changes in the intracellular distribution of Armadillo. *Development* **120**, 369–380 (1994).
26. Patel, N. H., Snow, P. M. & Goodman, C. S. Characterization and cloning of Fasciclin III: a glycoprotein expressed on a subset of neurons and axon pathways in *Drosophila*. *Cell* **48**, 975–988 (1987).
27. Berleth, T. *et al.* The role of localization of *bicoid* RNA in organizing the anterior pattern of the *Drosophila* embryo. *EMBO J.* **7**, 1749–1759 (1988).
28. Ephrussi, A., Dickinson, L. K. & Lehmann, R. *oskar* organizes the germ plasm and directs localization of the posterior determinant *nanos*. *Cell* **66**, 37–50 (1991).
29. Neumann-Silberberg, F. S. & Schüpach, T. The *Drosophila* dorsoventral patterning gene *gurken* produces a dorsally localized RNA and encodes a TGF α -like protein. *Cell* **75**, 165–174 (1993).
30. King, R. C. *Ovarian Development in Drosophila melanogaster* (Academic, New York, 1970).

Acknowledgements. We thank T. Uemura, M. Peifer, R. Lehmann and P. Lasko for providing reagents; the Developmental Studies Hybridoma Bank for the 7G10 antibody; H. Lipshitz and H. Krause for reading the manuscript; and R. Avancini for the *DE*-cadherin staining of *egl* mutant ovaries. This research was supported by a grant by the National Cancer Institute of Canada (to U.T.) with funds from the Terry Fox run.

Correspondence and request for materials should be addressed to D.G. (e-mail: dgodt@zoo.utoronto.ca) or U.T. (e-mail: utepass@zoo.utoronto.ca).

A function for lipoxygenase in programmed organelle degradation

Klaus van Leyen*, Robert M. Duvoisin†‡, Harald Engelhardt§ & Martin Wiedmann*‡

* Cellular Biochemistry and Biophysics Program, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, USA

† Dyson Vision Research Institute, Cornell University Medical College, and

‡ Cornell University Graduate School of Medical Sciences, 1300 York Avenue, New York, New York 10021, USA

§ Abteilung Molekulare Strukturbiologie, Max-Planck-Institut für Biochemie, D-82152 Martinsried, Germany

Membrane-enclosed organelles, a defining characteristic of eukaryotic cells, are lost during differentiation of specific cell types such as reticulocytes (an intermediate in differentiation of erythrocytes), central fibre cells of the eye lens, and keratinocytes¹. The degradation of these organelles must be tightly regulated with respect to both the time of activation and the specificity of membrane degradation. The expression of 15-lipoxygenase (15-LOX) peaks in reticulocytes immediately before organelle degradation². Here we show that 15-LOX integrates into the membranes of various organelles, allowing release of proteins from the organelle lumen and access of proteases to both luminal and integral membrane proteins. In addition, by sparing the plasma membrane, 15-LOX shows the required specificity for organelle membranes. Thus, the action of 15-LOX provides a mechanism by which the natural degradation process can be explained. This conclusion is supported by our finding that lipoxygenase expression in the eye lens is restricted to the region at which organelle degradation occurs.

Using arachidonate as substrate, mammalian lipoxygenases catalyse the biosynthesis of leukotrienes and lipoxins, which are important regulators of inflammation^{3,4}. More than 20 years ago, it was proposed that 15-LOX is involved in degrading mitochondria during terminal differentiation of erythrocytes⁵. Since then, most studies of 15-LOX have focused on its enzymatic activity, that is, oxygenation of membrane phospholipids^{2,6}. Here we demonstrate a new structural function for 15-LOX.

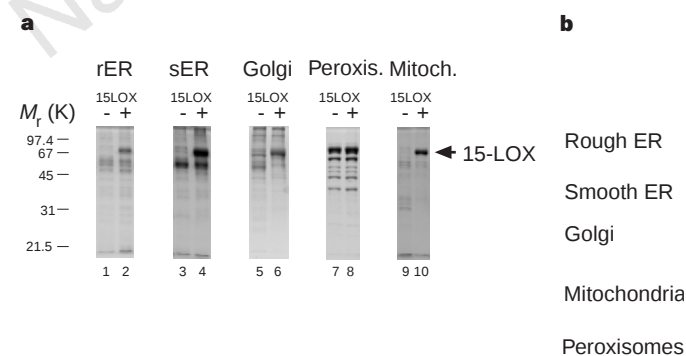


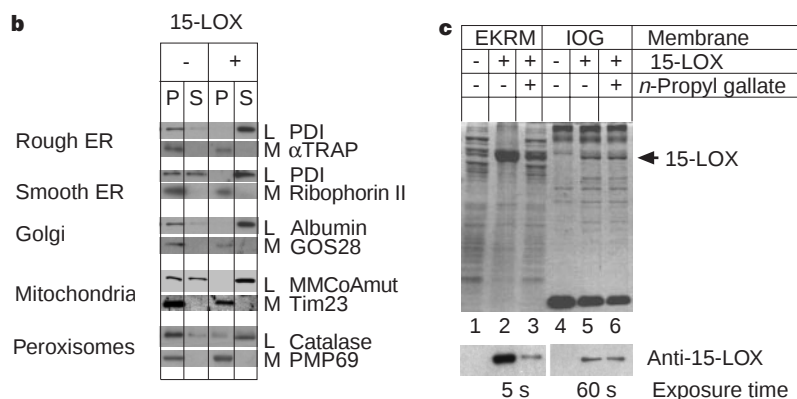
Figure 1 15-LOX binds to and permeabilizes intracellular organelles, but not the plasma membrane *in vitro*. **a**, 15-LOX associates with all organelles shown. Membrane preparations were incubated in the absence or presence of 15-LOX and re-isolated by sedimentation through a sucrose cushion. rER, rough ER; sER, smooth ER. **b**, Organelles are permeabilized by incubation with 15-LOX. After 2 h incubation and sedimentation through 0.5 M sucrose, pellet (P) and supernatant (S) fractions were immunoblotted for the indicated luminal (L) or integral membrane (M) proteins. Although we used 1 mg 15-LOX per ml here, 125 μ g ml⁻¹ is

To test the association of 15-LOX with organelles, we used sedimentation assays (Fig. 1a). We found that large amounts of 15-LOX bind to several organelles. Because untreated peroxisomes contained an endogenous protein that migrated at the same position as 15-LOX, we confirmed 15-LOX binding by immunoblotting (results not shown). Immunogold electron microscopy has recently been used to show that 15-LOX also associates with mitochondria in reticulocytes⁷.

We studied the dependence of 15-LOX binding on enzymatic activity by using the lipoxygenase inhibitors *n*-propyl gallate and eicosatetraynoic acid. These inhibitors block binding to endoplasmic reticulum (ER) membranes (Fig. 1c, lanes 1–3, and data not shown). Thus, 15-LOX enzymatic activity is required for binding. EGTA also blocked binding, but simultaneous addition of EGTA and calcium restored binding of 15-LOX to ER membranes (data not shown), which confirms earlier reports^{7,8} of a calcium requirement for membrane association.

We observed that the protein pattern from pelleted membranes changed after prolonged incubation with 15-LOX and speculated that this might be due to release of proteins from the lumen. Therefore, we determined whether organelles were permeabilized by immunoblotting for characteristic luminal proteins in the pellet and supernatant fractions after incubation for 2 h in the absence or presence of 15-LOX (Fig. 1b). To show that organelle membranes still pelleted and were not generally solubilized, we also probed integral membrane proteins by immunoblotting. Without exception, membrane proteins localized to the pellet whereas luminal proteins were released into the supernatant. However, both luminal proteins (GRP94, Bip, ERp72 and PDI) and integral membrane proteins (Sec61- α and ribophorin II) became accessible to protease digestion after treatment of ER membranes with 15-LOX (see Supplementary Information).

Organelle degradation in reticulocytes requires a high degree of specificity because the reticulocyte's plasma membrane is not degraded. To determine whether 15-LOX shows this specificity, we analysed 15-LOX binding either to ER membranes or to inside-out 'ghosts' made from erythrocytes, which expose the intracellular face of the plasma membrane towards the outside. After incubation, membranes were sedimented through a sucrose cushion and proteins were separated by SDS–polyacrylamide gel electrophoresis (SDS–PAGE) and either stained (Fig. 1c, top panel) or immunoblotted (Fig. 1c, bottom panel) for 15-LOX. Large amounts (~50%) of 15-LOX bound to ER membranes, but not to inside-out ghost membranes (Fig. 1c). Furthermore, we found that the inhibitor



sufficient for permeabilization of ER membranes (see Supplementary Information). **c**, 15-LOX binds inefficiently to erythrocyte membranes. EDTA-high-salt-stripped rough microsomes (ECRM) and inside-out erythrocyte ghosts (IOG) were incubated for 2 h at 26°C with either buffer, 15-LOX or 15-LOX pretreated with 15-LOX inhibitor (*n*-propyl gallate). After sedimentation through 0.5 M sucrose, proteins were separated by 12% SDS–PAGE and visualized by Coomassie staining (upper panel) or immunoblotting for 15-LOX (lower panels, with exposure times for ECL detection).

n-propyl gallate did not alter the weak binding of 15-LOX to inside-out ghosts, indicating a nonspecific association. The reduced binding of 15-LOX to plasma membranes may be due either to proteins found on the erythrocyte's cytoplasmic leaflet or to its higher cholesterol content compared with internal membranes (24% compared with 5–10%, respectively)⁹.

How does membrane binding by 15-LOX cause release of luminal proteins from organelles? It is clear that 15-LOX permeabilizes these membranes but does not solubilize them, as membranes can be sedimented after 15-LOX treatment. In our view, the most attractive model is that 15-LOX induces pore formation by integrating into membranes. To test this, we characterized the behaviour of 15-LOX by using Triton X-114 phase separation (Fig. 2a). After Triton X-114 extraction, integral membrane proteins partition into the detergent phase, whereas other proteins partition into the aqueous phase. As expected for a water-soluble protein, 15-LOX is found in the aqueous phase but, surprisingly, is recovered from the detergent phase after incubation with membranes. 15-LOX also behaves like an integral membrane protein after incubation with membranes and carbonate extraction (see Supplementary Information).

After incubation with membranes, a fraction of 15-LOX migrated more slowly during SDS-PAGE even after boiling in SDS sample buffer (Fig. 2b). This indicates that 15-LOX may form unusually stable oligomeric complexes after contact with membranes. To test whether oligomerization was induced by membranes indirectly or whether it required membrane contact, we preincubated 15-LOX with *n*-propyl gallate to inhibit membrane binding. This pretreatment also reduced the amount of more slowly migrating material. The oligomeric complexes were also obtained when samples were heated at 26°C or 40°C (instead of boiling) and when liposomes containing arachidonate as the acyl chain of the phospholipid were used (not shown). Consistent with these results, we also observed a much earlier elution of 15-LOX from sizing columns after detergent extraction from the membrane (not shown).

The association of 15-LOX with, and subsequent permeabilization of, membranes could be mediated by membrane lipids and/or proteins. Therefore we tested whether 15-LOX can permeabilize liposomes by preparing multilamellar liposomes from two distinct phosphatidylcholines²⁵. In contrast to other lipoxygenases, 15-LOX can act directly on phospholipids containing arachidonic acid. Therefore, 1-stearoyl-2-arachidonoyl-*sn*-glycero-3-phosphocholine is a potential substrate for 15-LOX, whereas 1-oleyl-2-palmitoyl-*sn*-

glycero-3-phosphocholine is not. Carboxyfluorescein, a water-soluble fluorescent dye, was incorporated into these liposomes (labelled as carboxyfluorescein-arachidonoyl liposomes and carboxyfluorescein-oleyl liposomes, respectively, in Fig. 3a). After incubation with buffer, aprotinin (as a control protein), 15-LOX or Triton X-100 (to solubilize the liposomes completely), the liposomes were sedimented by centrifugation and the amount of carboxyfluorescein released into the supernatant was measured. 15-LOX released carboxyfluorescein from carboxyfluorescein-arachidonoyl liposomes as efficiently as did Triton X-100, indicating that 15-LOX permeabilized liposomes. Thus, although they might facilitate the process, the presence of proteins in the membrane is not essential for membrane permeabilization.

Liposomes can still be sedimented and visualized by electron microscopy after 15-LOX treatment, indicating that membranes are not completely solubilized by 15-LOX, although we cannot exclude a partial solubilization. If 15-LOX-dependent permeabilization produces structures that are large enough to release proteins such as PDI and albumin (Fig. 1b), these structures should be visible by electron microscopy. Unilamellar liposomes incubated with 15-LOX and negatively stained with uranyl acetate revealed structures with diameters of 3–10 nm that were not present in controls (Fig. 3b). ER membranes incubated with 15-LOX produced similar structures (7–17 nm; results not shown). These diameters are sufficient to release proteins with relative molecular masses as large as 400,000 (400K).

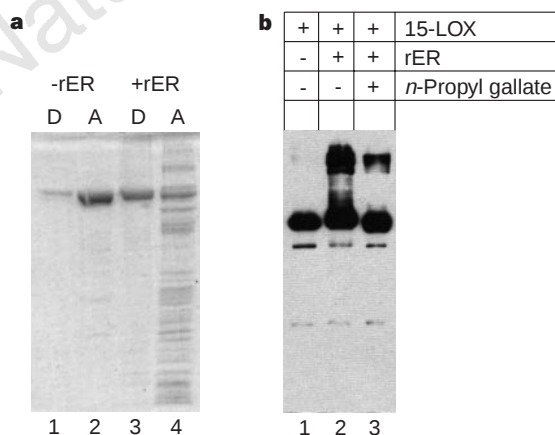


Figure 2 On incubation with rER membranes, 15-LOX converts from a soluble monomeric protein into an oligomeric membrane-bound structure. **a**, 15-LOX fractionates to the Triton X-114 detergent phase (D) if rER membranes are present; otherwise it remains in the aqueous phase (A). **b**, 15-LOX forms stable complexes of high relative molecular mass after membrane integration. 15-LOX was incubated with either buffer, or rER membranes, or was pretreated with the inhibitor *n*-propyl gallate and then incubated with rER membranes. The mixture was separated on 10% SDS-PAGE and immunoblotted for 15-LOX.

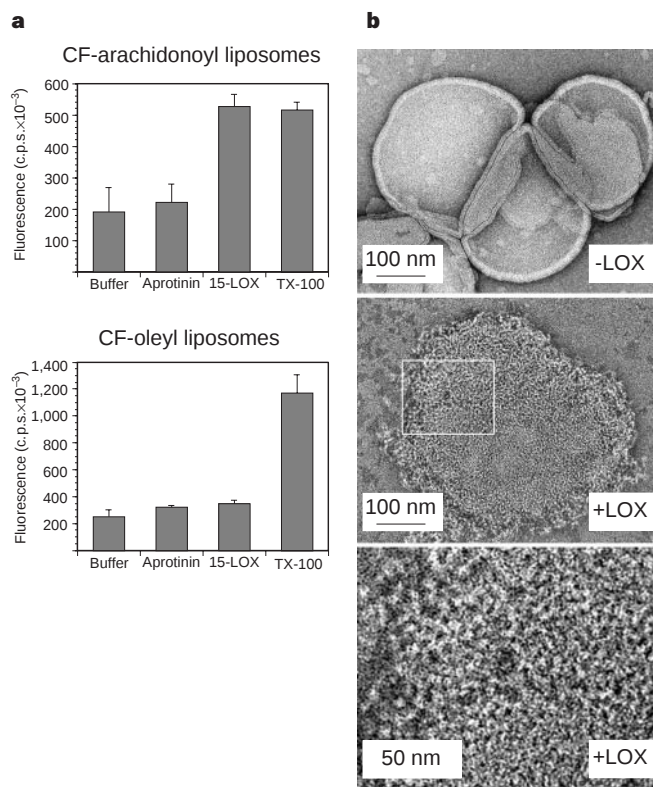


Figure 3 Membrane proteins are not required for membrane integration of 15-LOX. **a**, Release of carboxyfluorescein (CF) from liposomes after incubation with 15-LOX. Multilamellar vesicles containing carboxyfluorescein were generated from phosphatidylcholine with (upper panel) or without (lower panel) arachidonate as an acyl chain and incubated with buffer, aprotinin, 15-LOX or 1% Triton X-100. After re-sedimentation of liposomes, the supernatant's fluorescence emission was measured. **b**, Ultrastructural analysis of treated and untreated liposomes by electron microscopy. Unilamellar liposomes were incubated with buffer (upper panel) or 15-LOX (lower panels) and analysed by negative staining with uranyl acetate. Dark regions of the 15-LOX-treated liposomes represent stain-filled cavities and indicate pore-like structures and holes in the lipid membrane.

The reticulocyte is not the only cell type known to eliminate its organelles during terminal differentiation. Fibre cells of the eye lens are continuously generated from outer epithelial cells as the lens develops and expands. As fibre cells mature, organelles and chromatin are progressively degraded so that the cell becomes transparent to incoming light. A region in the centre of the lens distinguished by an organelle-free zone is surrounded by an intermediate stage, in which organelle degradation is ongoing. The involvement of lipoxygenase in this process has been suggested¹¹ but has not yet been demonstrated.

Double-labelling experiments were performed using affinity-purified rabbit antibodies against 15-LOX and a monoclonal mouse antibody that recognizes ER-resident proteins containing the KDEL Lys-Asp-Glu-Leu retrieval motif¹⁰. Indocarbocyanine (Cy3)-coupled anti-rabbit (Fig. 4a) and fluorescein isothiocyanate (FITC)-coupled anti-mouse (Fig. 4b) antibodies were used as secondary antibodies (red and green, respectively; Fig. 4c) to visualize lipoxygenase expression and the ER compartment. As expected, the anti-KDEL antibody gives intensive staining of the epithelial cell layer and adjacent cell layers, but the staining intensity gradually decreases towards the centre of the lens, indicating gradual breakdown of the ER. The organelle-free zone lacks ER staining, as reported previously¹².

There is strong lipoxygenase staining of areas in which organelle degradation occurs. Superposition of staining patterns (Fig. 4c) showed that a decrease in ER staining was accompanied by an increase in lipoxygenase staining, which is consistent with a function for lipoxygenase in organelle degradation in the lens. Epithelial cells in culture express a platelet-type lipoxygenase, a mediator of inflammation¹³. The epithelial cell layer was stained by the antibody that recognizes several lipoxygenases; adjacent cell layers do not contain detectable levels of lipoxygenase. Because epithelial cells contain normal levels of ER and other organelles, it is unlikely that the lipoxygenase staining here is indicative of organelle degradation. It has been proposed that nuclear degradation in the lens is an

abortive apoptotic pathway¹⁴: 15-LOX may act to mediate this process.

We have shown that bound 15-LOX integrates into organelle membranes and permeabilizes them. This activity of 15-LOX is reminiscent of the ability of several bacterial toxins, and of complement, to form pores. In contrast to the toxins, which attack the plasma membrane^{15,16}, 15-LOX is specific for organelle membranes. We propose that 15-LOX initiates organelle degradation by pore formation in all organelle membranes, allowing subsequent degradation, probably by the ubiquitin/proteasome system^{17,18}. The temporal pattern of lipoxygenase expression in reticulocytes is consistent with this role. Translation of 15-LOX is repressed by proteins that bind to the 3'-untranslated region of the messenger RNA¹⁹ until the nucleus is ejected, coinciding with organelle degradation. At this stage, cellular levels of 15-LOX can be as high as 4 mg ml⁻¹ (ref. 20). Thus, the action of 15-LOX on membranes is likely to direct the ever-present catabolic degradative machinery to appropriate sites. □

Methods

Purification of 15-LOX. 15-LOX was prepared from rabbit reticulocyte lysate²¹. After sedimentation of the ribosomes, the supernatant was precipitated with 66% ammonium sulphate and 15-LOX was purified on Q- and S-Sepharose. After ammonium sulphate precipitation and dialysis against 50 mM potassium acetate, 50 mM HEPES, pH 7.5, 5 mM magnesium acetate, 1 mM dithiothreitol (DTT), the enzyme was >95% pure. See also Supplementary Information.

Antibodies. Antibodies were raised in rabbits against the carboxy terminus of 15-LOX (sequence YLRPSIVENSVAI) and affinity-purified using the same peptide.

Binding and release assays. Organelles and antibodies were provided by G. Kreibich, W. W. Just, C. Hughes, W. A. Fenton, M. F. Bauer and M. Gmachl, or purchased from Stressgen (anti-PDI) and Sigma (anti-albumin). Sedimentation assays contained organelles at a final concentration of ~1 mg protein ml⁻¹, 50 mM HEPES/KOH, pH 7.5, 50 mM potassium acetate, 5 mM magnesium acetate, 1 mM DTT and 1 mg ml⁻¹ 15-LOX or buffer in 40 µl total volume. After incubation for 5 min at 26 °C and 10 min on ice, the membranes were sedimented through a 0.5 M sucrose cushion, resuspended in SDS sample buffer and separated by SDS-PAGE. Proteins were visualized by staining with Coomassie blue or western blotting with ECL detection (Amersham). After 2 h of incubation and centrifugation, aliquots of supernatant and pellet fractions were probed with antibodies against luminal and integral membrane proteins of the respective organelle. Inside-out ghosts were prepared from fresh rabbit blood²² and rough ER membranes and EDTA-high-salt-stripped rough microsomes were prepared from dog pancreas²³. Phase separation in Triton X-114 was done according to ref. 24.

Liposome preparation and electron microscopy. Multilamellar vesicles containing carboxyfluorescein (Molecular Probes) were generated from either 1-stearoyl-2-arachidonyl-*sn*-glycero-3-phosphocholine or 1-oleyl-2-palmitoyl-*sn*-glycero-3-phosphocholine (Avanti Polar Lipids)²⁵ and isolated by gel filtration on Sephadex G-25. After incubation with either buffer or aprotinin (as control) or with 15-LOX or 1% Triton X-100 (to achieve 100% release of carboxyfluorescein), liposomes were sedimented for 20 min at 45,000 r.p.m. in a TLA45 rotor (Beckman). Supernatants were adjusted to 1% Triton X-100 and fluorescence emission was measured at 515 nm (excitation wavelength 490 nm) with a Spex Fluorolog-2 spectrofluorometer.

For electron microscopy, we generated unilamellar vesicles²⁶. These vesicles were incubated with buffer or 0.85 mg ml⁻¹ 15-LOX for 30 min at 26 °C, sedimented for 20 min at 14,000 r.p.m. in an Eppendorf centrifuge, and prepared for negative staining with 2% uranyl acetate. Micrographs were taken in a Philips EM 420 electron microscope.

Lens immunohistochemistry. Adult mouse eyes that had been fixed in 4% paraformaldehyde overnight were paraffin-embedded and sectioned horizontally at 5 µm. Sections permeabilized with 0.075% saponin were blocked with 5% non-immune goat serum/0.2% albumin and stained with antibodies against 15-LOX and KDEL (Stressgen), respectively. Cy3-conjugated goat anti-rabbit and FITC-conjugated goat anti-mouse secondary antibodies (Jackson ImmunoResearch) were used for detection in a Zeiss Axiovert 35 microscope.

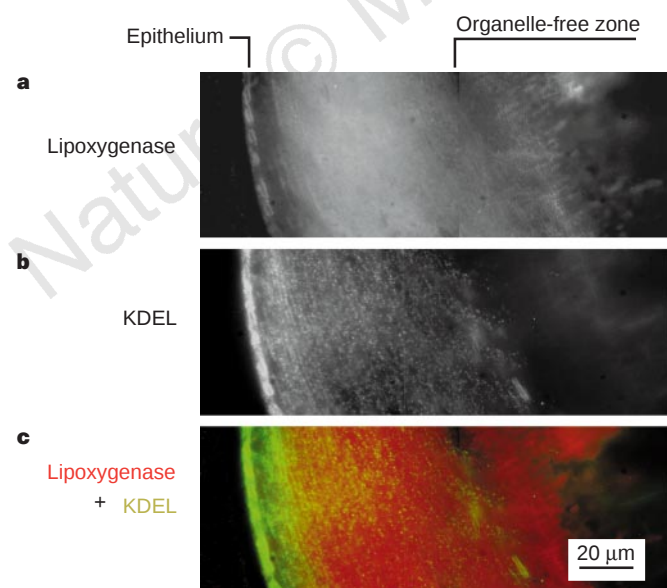


Figure 4 Lipoxigenase expression and organelle degradation in the adult mouse lens, indicating marked lipoxigenase expression in the inner cortex of the lens where organelles are degraded. **a**, Affinity-purified antiserum raised against lipoxigenase stains the outer part of the lens, including epithelium and outer fibre cells. An intermediate region between the epithelium and the outer fibre cells is, however, not stained. **b**, Organelles visualized using an antibody that recognizes KDEL proteins present in the endoplasmic reticulum. **c**, False colour superposition of **a** and **b**, showing the progressive degradation of organelles (decreasing green staining) coincident with the marked presence of lipoxigenase (red).

Received 21 May; accepted 21 July 1998.

- Alberts, B. et al. *Molecular Biology of the Cell* 1158 (Garland, New York, 1994).
- Schewe, T., Rapoport, S. M. & Kühn, H. Enzymology and physiology of reticulocyte lipoygenase: comparison with other lipoygenases. *Adv. Enzymol. Relat. Areas Mol. Biol.* **58**, 191–272 (1986).
- Samuelsson, B., Haeggstrom, J. Z. & Wetterholm, A. Leukotriene biosynthesis. *Ann. NY Acad. Sci.* **629**, 89–99 (1991).
- Kühn, H. & Chan, L. The role of 15-lipoxygenase in atherogenesis: pro- and antiatherogenic actions. *Curr. Opin. Lipidol.* **8**, 111–117 (1997).
- Schewe, T., Halang, W., Hiesch, C. & Rapoport, S. M. A lipoygenase in rabbit reticulocytes which attacks phospholipids and intact mitochondria. *FEBS Lett.* **60**, 149–152 (1975).
- Kühn, H., Belkner, J., Wiesner, R. & Brash, A. R. Oxygenation of biological membranes by the pure reticulocyte lipoygenase. *J. Biol. Chem.* **265**, 18351–18361 (1990).
- Brinckmann, R. et al. Membrane translocation of 15-lipoxygenase in hematopoietic cells is calcium-dependent and activates the oxygenase activity of the enzyme. *Blood* **91**, 64–74 (1998).
- Watson, A. & Doherty, F. J. Calcium promotes membrane association of reticulocyte 15-lipoxygenase. *Biochem. J.* **298**, 377–383 (1994).
- Jain, M. K. & Wagner, R. C. *Introduction to Biological Membranes* (Wiley, New York, 1980).
- Munro, S. & Pelham, H. R. A C-terminal signal prevents secretion of luminal ER proteins. *Cell* **48**, 899–907 (1987).
- Hunt, T. On the translational control of suicide in red cell development. *Trends Biochem. Sci.* **14**, 393–394 (1989).
- Bassnett, S. & Mataic, D. Chromatin degradation in differentiating fiber cells of the eye lens. *J. Cell Biol.* **137**, 37–49 (1997).
- Arora, J. K., Lysz, T. W. & Zelenka, P. S. A role for 12(s)-hete in the response of human lens epithelial cells to epidermal growth factor and insulin. *Invest. Ophthalmol. Vis. Sci.* **37**, 1411–1418 (1996).
- Ishizaki, Y., Jacobson, M. D. & Raff, M. C. A role for caspases in lens fiber differentiation. *J. Cell Biol.* **140**, 153–158 (1998).
- Gouaux, E. Channel-forming toxins: tales of transformation. *Curr. Opin. Struct. Biol.* **7**, 566–573 (1997).
- Palmer, M. et al. Assembly mechanism of the oligomeric streptolysin O pore: the early membrane lesion is lined by a free edge of the lipid membrane and is extended gradually during oligomerization. *EMBO J.* **17**, 1598–1605 (1998).
- Mayer, A., Siegel, N. R., Schwartz, A. L. & Ciechanover, A. Degradation of proteins with acetylated amino termini by the ubiquitin system. *Science* **244**, 1480–1483 (1989).
- Coux, O., Tanaka, K. & Goldberg, A. L. Structure and functions of the 20S and 26S proteasomes. *Annu. Rev. Biochem.* **65**, 801–847 (1996).
- Ostareck, D. H. et al. mRNA silencing in erythroid differentiation: hnRNP K and hnRNP E1 regulate 15-lipoxygenase translation from the 3' end. *Cell* **89**, 597–606 (1997).
- Rapoport, S. M. et al. The lipoygenase of reticulocytes. Purification, characterization and biological dynamics of the lipoygenase; its identity with the respiratory inhibitors of the reticulocyte. *Eur. J. Biochem.* **96**, 545–561 (1979).
- Jackson, R. J. & Hunt, T. Preparation and use of nuclease-treated rabbit reticulocyte lysates for the translation of eukaryotic messenger RNA. *Methods Enzymol.* **96**, 50–74 (1983).
- Steck, T. L. & Kant, J. A. Preparation of impermeable ghosts and inside-out vesicles from human erythrocyte membranes. *Methods Enzymol.* **31A**, 172–180 (1974).
- Walter, P. & Blobel, G. Preparation of microsomal membranes for cotranslational protein translocation. *Methods Enzymol.* **96**, 84–93 (1983).
- Bordier, C. Phase separation of integral membrane proteins in Triton X-114 solution. *J. Biol. Chem.* **256**, 1604–1607 (1981).
- New, R. R. C. *Liposomes: A Practical Approach* 37–39 (Oxford Univ. Press, New York, 1990).
- Paul, A., Engelhardt, H., Jakubowski, U. & Baumeister, W. Two-dimensional crystallization of a bacterial surface protein on lipid vesicles under controlled conditions. *Biophys. J.* **61**, 172–188 (1992).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements. We thank D. Smith and J. Olesker for assistance in writing the manuscript; T. Söllner, J. E. Rothman, P. Szabo, G. van Meer, D. Nikolov, A. Koff, G. Bacher, and members of the Wiedmann, Duvoisin and Rothman laboratories for discussions and comments; N. Min and L. Cohen-Gould for performing the initial immunohistochemistry and electron microscopy experiments. We thank P. Marks and R. Rifkind for suggesting the eye lens experiments. This work was supported by the Memorial Sloan-Kettering Cancer Center, a Fellowship by the Deutsche Forschungsgemeinschaft (to K.v.L.), and by the Samuel and May Rudin Foundation and a Tolly Vinik Pilot Grant Award (to R.M.D.).

Correspondence and requests for materials should be addressed to M.W.

A protein conjugation system essential for autophagy

Noboru Mizushima*, Takeshi Noda*, Tamotsu Yoshimori*, Yae Tanaka†, Tomoko Ishii†, Michael D. George‡, Daniel J. Klionsky‡, Mariko Ohsumi† & Yoshinori Ohsumi*

* Department of Cell Biology, National Institute for Basic Biology, Okazaki 444-8585, Japan

† Department of Biosciences, Teikyo University of Science & Technology, Yamanashi 409-0193, Japan

‡ Section of Microbiology, University of California, Davis, California 95616, USA

Autophagy is a process for the bulk degradation of proteins, in which cytoplasmic components of the cell are enclosed by double-membrane structures known as autophagosomes for delivery to lysosomes or vacuoles for degradation^{1–4}. This process is crucial for survival during starvation and cell differentiation. No molecules have been identified that are involved in autophagy in

higher eukaryotes. We have isolated 14 autophagy-defective (*apg*) mutants of the yeast *Saccharomyces cerevisiae*⁵ and examined the autophagic process at the molecular level^{6–9}. We show here that a unique covalent-modification system is essential for autophagy to occur. The carboxy-terminal glycine residue of Apg12, a 186-amino-acid protein, is conjugated to a lysine at residue 149 of Apg5, a 294-amino-acid protein. Of the *apg* mutants, we found that *apg7* and *apg10* were unable to form an Apg5/Apg12 conjugate. By cloning *APG7*, we discovered that Apg7 is a ubiquitin-E1-like enzyme. This conjugation can be reconstituted *in vitro* and depends on ATP. To our knowledge, this is the first report of a protein unrelated to ubiquitin that uses a ubiquitination-like conjugation system. Furthermore, Apg5 and Apg12 have mammalian homologues, suggesting that this new modification system is conserved from yeast to mammalian cells.

In yeast, autophagy is induced by various starvation conditions, and its progression is easily monitored under a light microscope¹: when wild-type cells were cultured under nitrogen-starvation conditions in the presence of phenylmethylsulphonyl fluoride (PMSF), autophagic bodies accumulated in the vacuoles (arrows in Fig. 1a). The *apg12-1* mutant did not accumulate autophagic bodies during starvation. We cloned the *APG12* gene by the method described previously^{7,8}. *APG12* encodes a hydrophilic protein of 186 amino acids with a predicted relative molecular mass (M_r) of 21K (Fig. 1b).

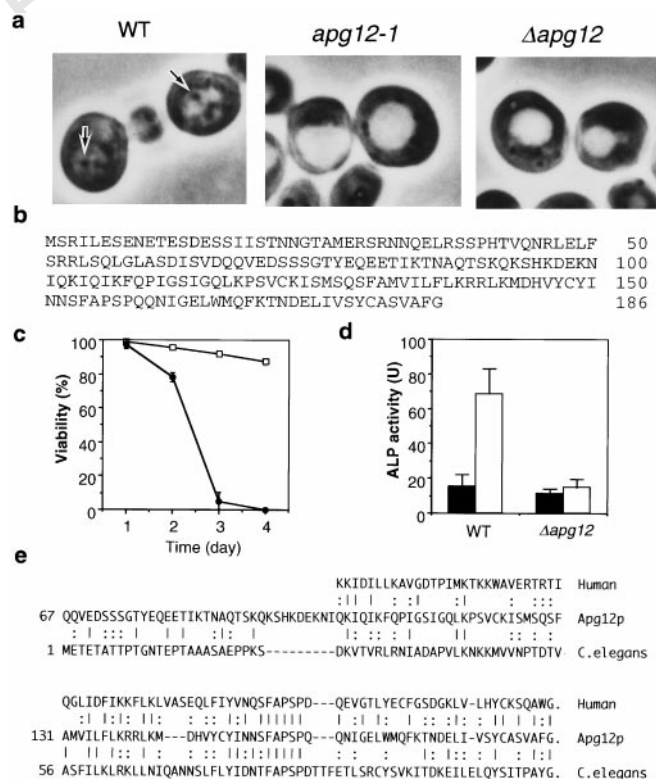


Figure 1 Cloning of *APG12* and phenotype of *apg12* disruptant. **a**, Wild-type, *apg12-1* mutant and Δ *apg12* cells were cultured in nitrogen-starvation medium containing 1 mM PMSF. After incubation for 6 h, cells were observed under a phase-contrast microscope. Arrows indicate autophagic bodies. **b**, Amino-acid sequence of Apg12. **c**, Wild-type (squares) and Δ *apg12* (circles) were cultured in nitrogen-starvation medium and their viability was determined by phloxine B staining⁵. **d**, Quantification of autophagic activity of wild-type and Δ *apg12* cells by alkaline phosphatase (ALP) assay before (black bars) and after (white bars) nitrogen starvation for 4 h. Error bars indicate s.d. of three independent experiments. **e**, Homology between Apg12 and potential human and *C. elegans* counterparts. *C. elegans* U32305 is 46% similar and 22% identical to amino acids 67–186 of yeast Apg12. A human cDNA (THC173313) encodes a protein that is 59% similar and 32% identical to amino acids 102–186 of Apg12.

Gene disruption experiments revealed that *APG12* is not essential for growth (data not shown) but is essential for autophagy (Fig. 1a) and for maintaining viability during starvation (Fig. 1c). We confirmed this in an assay system for measuring autophagic activity (Fig. 1d), in which a truncated form of pro-alkaline phosphatase expressed in the cytoplasm was delivered to vacuoles in an autophagy-dependent manner and processed to the active enzyme¹⁰. A vacuolar enzyme, aminopeptidase I, is delivered from the cytoplasm to vacuoles constitutively to yield the mature, active enzyme¹¹. This 'Cvt pathway' is closely linked to the autophagic process¹², and all *apg* mutants¹³, including Δ *apg12* cells, show defects in this pathway (see Fig. 3d). The amino-acid sequence of Apg12 did not provide any insight into its function, but a BLAST search identified a potential *Caenorhabditis elegans* homologue whose function is unknown (Fig. 1e). In addition, a search of the EST (expressed-sequence tag) database identified several cDNA fragments encoding parts of a potential human homologue (Fig. 1e).

To detect Apg12, we constructed a 3 × haemagglutinin(HA)-tagged APG12. On immunoblotting, Apg12 presented as a ladder of bands between 31K–32.5K (Fig. 2a). As phosphatase treatment of the lysate yielded a single band at 31K representing tagged Apg12 (data not shown), we concluded that Apg12 is phosphorylated *in vivo*. Furthermore, we found that about half of the Apg12 was present as a much larger band of ~70K (asterisked in Fig. 2a, b). Although the 31K Apg12 was detected in all *apg* mutant strains, the Δ *apg5*, *apg7-1* and *apg10-1* strains did not show the 70K band (Fig. 2b; Δ *apg1* is representative of the other mutants). These results indicate that these three *apg* products are essential for the generation of the 70K band.

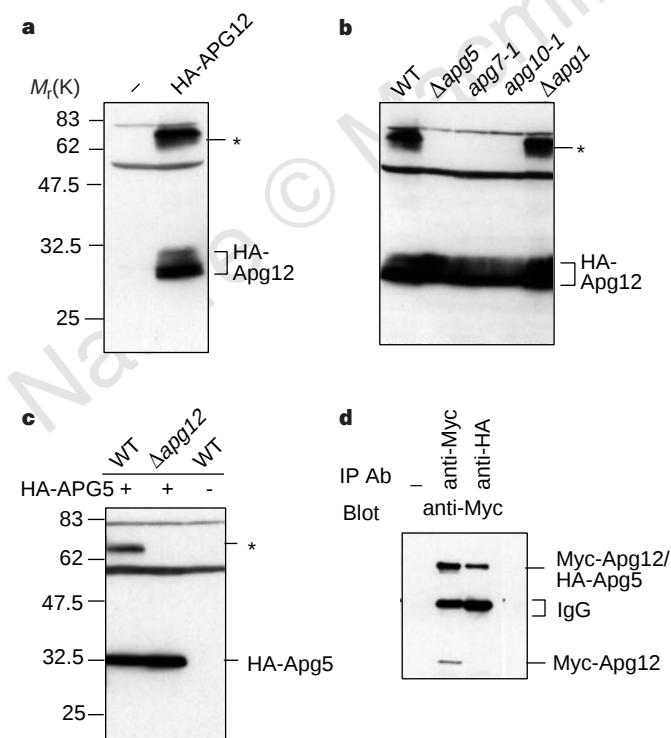


Figure 2 Apg12 is conjugated to Apg5. **a**, **b**, Lysates from Δ *apg12* cells carrying only vector or 3 × HA-APG12 (**a**), and Δ *apg5*, *apg7-1*, *apg10-1* and Δ *apg1* cells carrying 3 × HA-APG12 plasmid (**b**) were immunoblotted using anti-HA antibody. The positions of 3 × HA-Apg12 and the larger product (asterisks) are indicated. **c**, Immunoblot analysis of wild-type and Δ *apg12* cells harbouring HA-APG5 plasmid. **d**, Δ *apg5* Δ *apg12* cells were co-transformed with Myc-APG12 and HA-APG5. Their lysates were immunoprecipitated with anti-Myc or anti-HA antibodies and detected by immunoblotting using anti-Myc antibody. The position of the crossreacting IgG heavy chain is indicated.

We have previously shown that the *APG5* gene encodes a 294-amino-acid protein⁶. Immunoblot analysis of 1 × HA-tagged Apg5 indicated that it also generated two bands in nearly equal amounts, one of the size of tagged Apg5 (32.5K) and the other at about 70K (Fig. 2c). In the Δ *apg12* strain, the higher band was not seen, whereas the 32.5K band of Apg5 was slightly increased (Fig. 2c). Immunoprecipitation analysis revealed that the 70K band included both Apg5 and Apg12 (Fig. 2d). We concluded that it was a one-to-one conjugate of Apg5 and Apg12.

To characterize the 70K band further, we did mutagenic analysis of Apg12 (Fig. 3a). We found that the carboxy-terminal portion of Apg12 was important for the conjugation (Fig. 3b: Δ 57 and Δ 121).

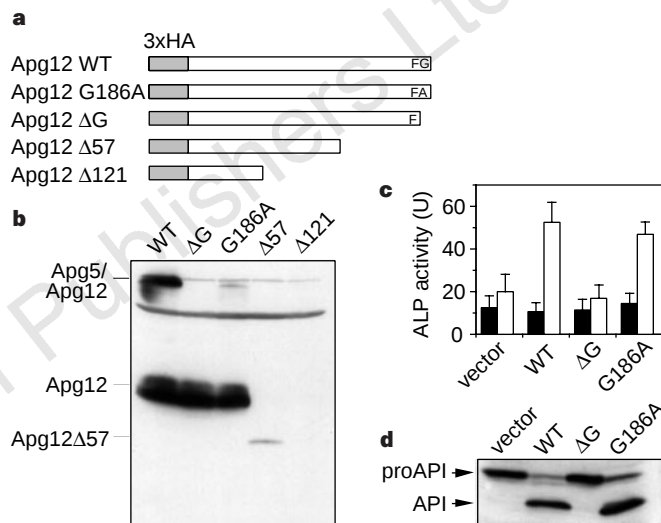


Figure 3 The C-terminal Gly residue of Apg12 is essential for interaction with Apg5 and for autophagy. **a**, Diagram of Apg12 C-terminal mutants. **b**, Δ *apg12* cells were transformed with the mutant plasmids and their lysates were immunoblotted with anti-HA antibody. **c**, Autophagic activity was measured as described for Fig. 1d. **d**, Transport of pro-API to the vacuole was examined by immunoblotting with anti-API antiserum. The positions of pro-API and mature API are indicated.

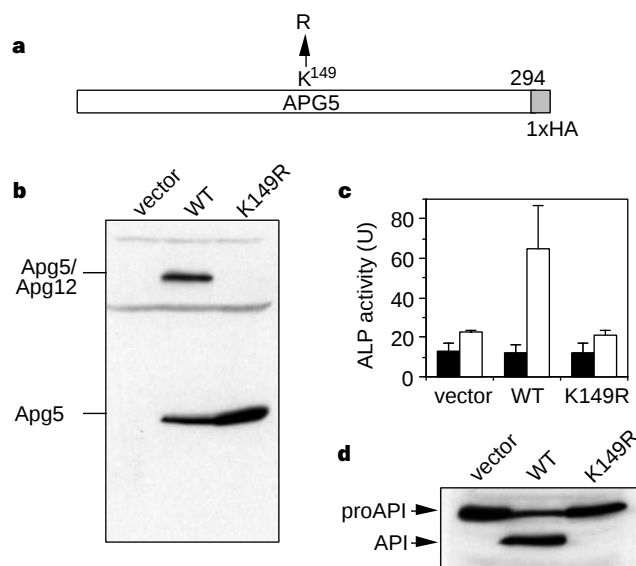


Figure 4 Apg5^{K149R} is unable to generate Apg5/Apg12 conjugate and is defective in autophagy. **a**, Position of the putative Apg12-interacting Lys residue. Δ *apg5* cells were transformed with vector alone, wild-type APG5 or APG5^{K149R}, and then immunoblotted with anti-HA (**b**) and anti-API (**d**). Autophagic activity was determined by alkaline phosphatase assay (**c**).

Even a single Gly 186 deletion at the C terminus (Apg12ΔG) caused complete loss of the Apg12/Apg5 conjugate, although free Apg12ΔG was detected in an amount comparable to that in the wild type (Fig. 3b: ΔG). Apg12^{G186A}, in which the Gly 186 is replaced by alanine, was incorporated into the higher band inefficiently, but still significantly (Fig. 3b: G186A). This indicates that Gly 186 is important for Apg5/Apg12 conjugation. We next assessed the functional activities of these mutants. Apg12ΔG showed an Apg-negative phenotype (Fig. 3c), and was also unable to produce mature aminopeptidase I (Fig. 3d), indicating that the Apg5/Apg12 conjugate is required for both autophagy and cytosol-to-vacuole targeting of this enzyme. The Apg12^{G186A} mutant showed an almost normal phenotype for autophagy and for maturation of aminopeptidase I (Fig. 3c, d), suggesting that a small amount of Apg5/Apg12 conjugate is enough for it to function normally.

By analogy with ubiquitin^{14–16}, conjugation of Apg5 and Apg12 probably occurs through formation of an isopeptide bond between the C-terminal Gly 186 of Apg12 and an ε-amino group of one of the 19 lysine residues in Apg5. To test this, we systematically replaced each lysine residue of Apg5 with arginine. Both free Apg5 and the Apg5/Apg12 conjugate were detected in 18 mutants (data not shown). The Apg5^{K149R} variant had no conjugate at all, but a higher amount of free Apg5^{K149R} (Fig. 4a, b), indicating that the Lys 149 residue of Apg5 is the acceptor site for Apg12 conjugation. As expected, Apg5^{K149R} was defective in both autophagy and in generating mature aminopeptidase I (Fig. 4c, d), whereas the other 18 mutants were normal (data not shown). Starvation did not alter the relative amounts of free Apg5, free Apg12 or of the Apg5/Apg12 conjugate. We conclude that the conjugate functions as a common machinery in both pathways: for the autophagic pathway during starvation and for the Cvt pathway in the growing phase.

As shown in Fig. 2, the *apg7* and *apg10* mutants failed to conjugate Apg5 and Apg12, suggesting that these two APG products

may function as an enzyme system for conjugation. Cloning of the *APG7* gene revealed that it encodes a 630-amino-acid protein with predicted *M_r* of 71.4K (Fig. 5a). The region containing amino acids 322–392 of Apg7 shows significant homology with the corresponding region in E1, the ubiquitin-activating enzyme in *S. cerevisiae* (Fig. 5b) and in other species (data not shown). This region encompasses a putative ATP-binding site (GxGxxG)¹⁷, suggesting that Apg7 may be an Apg12-activating enzyme. Although the sequence around the active-site cysteine is less conserved, alignments between Apg7 and other E1-like enzymes indicate that Cys 507 is a putative active-site cysteine (Fig. 5b). Apg10 might be an E2 ubiquitin-conjugating enzyme type of protein because its size is similar to various E2 enzymes and one of its cysteine residues is essential for its function (T. Shintani *et al.*, unpublished results). We reconstituted the conjugation reaction *in vitro*. Lysates of Δ*apg5* cells and Δ*apg12* cells were mixed *in vitro* and incubated with or without ATP. Figure 5c shows that the 70K band appeared in a time-dependent and ATP-dependent manner. The conjugation was sensitive to 1 mM *N*-ethylmaleimide (data not shown). These results show that the Apg12 conjugation pathway contains an ATP-dependent step, which is probably the activation of Apg12 by Apg7.

Autophagy involves a dynamic membrane rearrangement^{2–4}. Morphological studies have indicated that all APG products function before the autophagosome formation step (M. Baba and T.O., unpublished results). Some Apg proteins are present on membrane structures⁹. Most of the Apg5 and Apg5/Apg12 conjugate, and more than half of the free Apg12, were present in 100,000g pellet fractions (data not shown), suggesting that they associate with some membrane compartments. We therefore examined their intracellular localization by sucrose density-gradient centrifugation analysis and found that free Apg5 and the Apg5/Apg12 conjugate co-fractionated (Fig. 6); in contrast, most of the

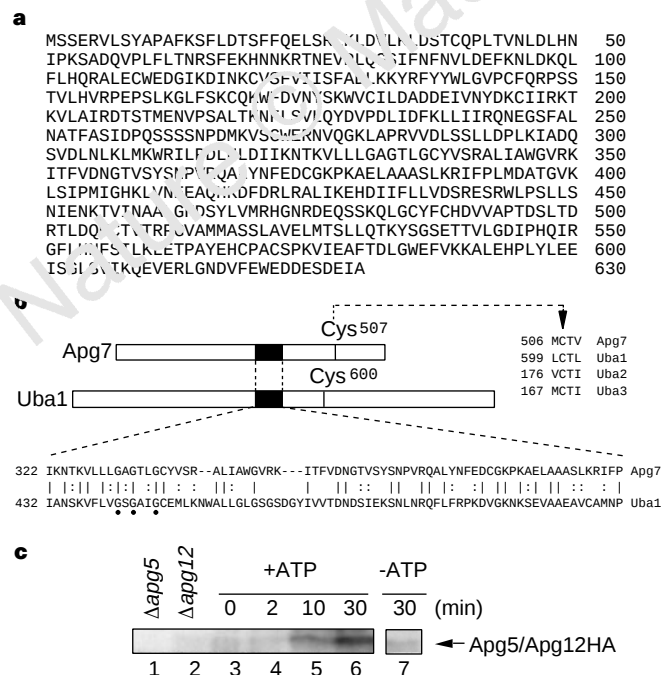


Figure 5 Apg7 is an E1-like protein and Apg12 is conjugated to Apg5 in an ATP-dependent manner. **a**, Amino-acid sequence of Apg7. **b**, Homology between Apg7 and Uba1, *S. cerevisiae* E1 enzyme. Black circles indicate a putative ATP-binding site (GxGxxG). The putative active-site Cys residue of Apg7 is indicated. **c**, *In vitro* conjugation of Apg5 and Apg12. A cell lysate of Δ*apg12* carrying HA-APG5 (lane 1) was incubated with an equal amount of lysate from Δ*apg5* carrying HA-APG12 (2 μg plasmid) (lane 2) at 30 °C with (lane 4–6) or without (lane 7) 5 mM ATP. Samples were mixed with SDS-sample buffer at the times indicated.

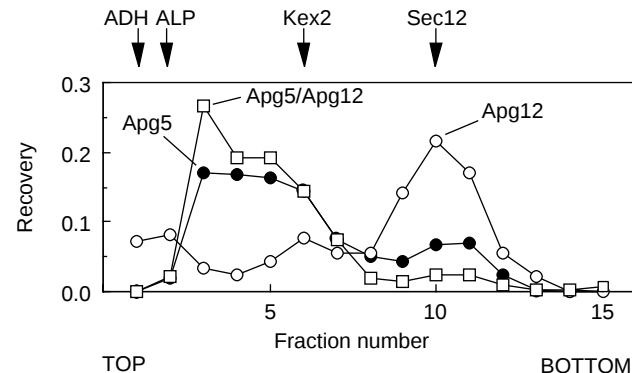


Figure 6 Apg5/Apg12 conjugate co-fractionates with free Apg5 but not free Apg12. Spheroplasts were generated from cells expressing either HA-Apg12 or HA-Apg5. Their lysates were mixed and layered on top of a 10-step (18–54% w/w) sucrose gradient, and centrifuged at 174,000g for 2.5 h (ref. 29). Fifteen fractions were collected and the positions of free Apg5, free Apg12 and Apg5/Apg12 conjugate were examined by western blotting. The peak fractions of alcohol dehydrogenase (ADH)(cytosol), ALP (vacuole), Kex2 (Golgi) and Sec12 (endoplasmic reticulum) are indicated by arrows.

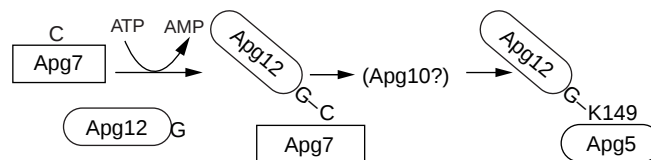


Figure 7 Model of the Apg12-conjugation system.

Apg12 was in the denser fractions. These results indicate that the conjugation of Apg5 and Apg12 is associated with a change in the subcellular localization of Apg12.

We have described a new covalent modification system that is required for autophagy in yeast. Four of 14 APG products function in this pathway. Our model is shown in Fig. 7: Apg12 is activated by binding to Apg7 via a high-energy thioester bond; through transfer to an E2-like molecule (possibly Apg10), Apg12 is finally conjugated to Lys 149 of Apg5 via an isopeptide bond. Although the steps in this conjugation pathway are similar to those that occur in ubiquitination^{14–16} and in the modification by other ubiquitin-like proteins such as SUMO-1 (refs 18–21), Smt3 (ref. 22), Rub1 (refs 23, 24) and Nedd8 (ref. 25), Apg12 has several unique features. It has no significant homology to ubiquitin and is much larger than ubiquitin and ubiquitin-related modifiers^{18–25}. Only a single specific substrate, Apg5, has been found. Apg12 homologues in human and *C. elegans* have a glycine residue at the C terminus (Fig. 1c). We have cloned human Apg12 and found that it is conjugated to human Apg5 (N.M., H. Sugita, T.Y. and Y.O., manuscript in preparation). Human Apg5 was recently cloned as 'apoptosis specific protein' by another group²⁶, although its physiological significance is not clear yet. This conjugation system is conserved from yeast to mammalian cells, and may be critical for autophagy in every eukaryote. □

Methods

Yeast strains. The *Saccharomyces cerevisiae* strains used for cloning and immunochemical analysis were MT3-4-4(MATa *apg12-1 ura3*), MT87-4-5(MATa *apg7-1 ura3*), MT91-4-2(MATa *apg10-1 ura3*), SKD5-1D(MATa *ura3 leu2 trp1 Δapg5::LEU2*) and YYK36(MATa *ura3 leu2 trp1 his3 Δapg1::LEU2*). Gene disruptions of APG5 and APG12 were performed with YW5-1B(MATa *ura3 leu2 trp1*) or KA31(MATa *ura3 leu2 trp1 his3*).

Alkaline phosphatase assay. The APG12 or APG5 gene was disrupted in TN125(MATa *ura3 leu2 trp1 his3 ade2 lys2 PHO8::pho8Δ60*), and the assay was done as described²⁷.

Immunochemical procedures. Whole-cell extracts were prepared by suspending cells in 0.2 M NaOH, 0.5% β-mercaptoethanol, and precipitated with acetone. Extracts were separated by SDS-PAGE, followed by immunoblotting using anti-HA antibody (16B12, BAbCO) or anti-API (aminopeptidase I) polyclonal antibody. Immunoprecipitation was done as described²⁸ using 16B12 or anti-Myc antibody (9E10).

Site-directed mutagenesis. Mutation and deletion constructs were generated by PCR-based site-directed mutagenesis and confirmed by automated DNA sequencing.

In vitro Apg12 conjugation assay. Total cell lysates were prepared from *Δapg12* strain expressing HA-Apg5 and *Δapg5* strain expressing HA-Apg12 after spheroplasting. Both lysates (30 mg ml⁻¹) were mixed in 50 mM Tris (pH 7.5), 100 mM NaCl, 10 mM MgCl₂, 1 mM DTT, 0.3 mM PMSF and 2 μg ml⁻¹ pepstatin, and incubated at 30 °C for the indicated times with or without 5 mM ATP. The reaction was stopped by mixing with SDS-PAGE buffer and boiling.

Received 20 May; accepted 29 June 1998.

1. Takeshige, K., Baba, M., Tsuboi, S., Noda, T. & Ohsumi, Y. Autophagy in yeast demonstrated with proteinase-deficient mutants and conditions for its induction. *J. Cell Biol.* **119**, 301–311 (1992).
2. Baba, M., Takeshige, K., Baba, N. & Ohsumi, Y. Ultrastructural analysis of the autophagic process in yeast: Detection of autophagosomes and their characterization. *J. Cell Biol.* **124**, 903–913 (1994).
3. Seglen, P. O. & Bohley, P. Autophagy and other vacuolar protein degradation mechanisms. *Experientia* **48**, 158–172 (1992).
4. Dunn, W. A. Jr Autophagy and related mechanisms of lysosome-mediated protein degradation. *Trends Cell Biol.* **4**, 139–143 (1994).
5. Tsukada, M. & Ohsumi, Y. Isolation and characterization of autophagy-defective mutants of *Saccharomyces cerevisiae*. *FEBS Lett.* **333**, 169–174 (1993).
6. Kametaka, S., Matsuura, A., Wada, Y. & Ohsumi, Y. Structural and functional analyses of APG5, a gene involved in autophagy in yeast. *Gene* **178**, 139–143 (1996).
7. Matsuura, A., Tsukada, M., Wada, Y. & Ohsumi, Y. Apg1p, a novel protein kinase required for the autophagic process in *Saccharomyces cerevisiae*. *Gene* **192**, 245–250 (1997).
8. Funakoshi, T., Matsuura, A., Noda, T. & Ohsumi, Y. Analyses of APG13 gene involved in autophagy in yeast, *Saccharomyces cerevisiae*. *Gene* **192**, 207–213 (1997).
9. Kametaka, S., Okano, T., Ohsumi, M. & Ohsumi, Y. Apg14p and Apg6p/Vps30p form a protein complex essential for autophagy in the yeast *Saccharomyces cerevisiae*. *J. Biol. Chem.* (in press).
10. Noda, T., Matsuura, A., Wada, Y. & Ohsumi, Y. Novel system for monitoring autophagy in the yeast *Saccharomyces cerevisiae*. *Biochem. Biophys. Res. Commun.* **210**, 126–132 (1995).
11. Klionsky, D. J., Cuevas, R. & Yaver, D. S. Aminopeptidase I of *Saccharomyces cerevisiae* is localized to the vacuole independent of the secretory pathway. *J. Cell Biol.* **119**, 287–299 (1992).

12. Baba, M., Osumi, M., Scott, S. V., Klionsky, D. J. & Ohsumi, Y. Two distinct pathways for targeting proteins from the cytoplasm to the vacuole/lysosome. *J. Cell Biol.* **139**, 1687–1695 (1997).
13. Scott, S. V. *et al.* Cytoplasm-to-vacuole targeting and autophagy employ the same machinery to deliver proteins to the yeast vacuole. *Proc. Natl. Acad. Sci. USA* **93**, 12304–12308 (1996).
14. Hochstrasser, M. Ubiquitin-dependent protein degradation. *Annu. Rev. Genet.* **30**, 405–429 (1996).
15. Varshavsky, A. The ubiquitin system. *Trends Biochem. Sci.* **22**, 383–387 (1997).
16. Weissman, A. M. Regulating protein degradation by ubiquitination. *Immunol. Today* **18**, 189–198 (1997).
17. McGrath, J. P., Jentsch, S. & Varshavsky, A. *UBA1*: an essential yeast gene encoding ubiquitin-activating enzyme. *EMBO J.* **10**, 227–236 (1991).
18. Matunis, M. J., Coutavas, E. & Blobel, G. A novel ubiquitin-like modification modulates the partitioning of the Ran-GTPase-activating protein RanGAP1 between the cytosol and the nuclear pore complex. *J. Cell Biol.* **135**, 1457–1470 (1996).
19. Mahajan, R., Delphin, C., Guan, T., Gerace, L. & Melchior, F. A small ubiquitin-related polypeptide involved in targeting RasGAP1 to nuclear pore complex protein RanBP2. *Cell* **88**, 97–107 (1997).
20. Mahajan, R., Gerace, L. & Melchior, F. Molecular characterization of the SUMO-1 modification of RanGAP1 and its role in nuclear envelope association. *J. Cell Biol.* **140**, 259–270 (1998).
21. Matunis, M. J., Wu, J. & Blobel, G. SUMO-1 modification and its role in targeting the Ran GTPase-activating protein, RanGAP1, to the nuclear pore complex. *J. Cell Biol.* **140**, 499–509 (1998).
22. Johnson, E. S., Schwenhorst, I., Dohmen, R. J. & Blobel, G. The ubiquitin-like protein Smt3p is activated for conjugation to other proteins by an Aosl1p/Uba2p heterodimer. *EMBO J.* **16**, 5509–5519 (1997).
23. Liakopoulos, D., Doenges, G., Matuschewski, K. & Jentsch, S. A novel protein modification pathway related to the ubiquitin system. *EMBO J.* **17**, 2208–2214 (1998).
24. Lammer, D. *et al.* Modification of yeast Cdc53p by the ubiquitin-related protein Rub1p affects function of the SCF^{Cdc4} complex. *Genes Dev.* **12**, 914–926 (1998).
25. Osaka, F. *et al.* A new Nedd8-ligating system for cullin-4A. *Genes Dev.* (in press).
26. Hammond, E. M. *et al.* Homology between a human apoptosis specific protein and the product of APG5, a gene involved in autophagy in yeast. *FEBS Lett.* **425**, 391–395 (1998).
27. Noda, T. & Ohsumi, Y. Tor, a phosphatidylinositol kinase homologue, controls autophagy in yeast. *J. Biol. Chem.* **273**, 3963–3966 (1998).
28. Kamada, Y. *et al.* Activation of yeast protein kinase C by Rho1 GTPase. *J. Biol. Chem.* **271**, 9193–9196 (1996).
29. Antebi, A. & Fink, G. R. The yeast Ca²⁺-ATPase homologue, PMR1, is required for normal Golgi function and localizes in a novel Golgi-like distribution. *Mol. Biol. Cell* **3**, 633–654 (1992).

Acknowledgements. We thank Y. Wada for the genomic library. N.M. is a research fellow of the Japan Society for the Promotion of Science.

Correspondence and requests for materials should be addressed to Y.O. (e-mail: yohsumi@nibb.ac.jp). The sequences of APG12 and APG7 are available from GenBank under accession numbers Z36086 (ORF YBR217w) and U00027 (ORF YHR171w), respectively.

Retinoid-X receptor signalling in the developing spinal cord

Ludmila Solomin*, Clas B. Johansson†, Rolf H. Zetterström‡, Reid P. Bissonnette§, Richard A. Heyman§, Lars Olson‡, Urban Lendahl†, Jonas Frisén† & Thomas Perlmann*

* The Ludwig Institute for Cancer Research, Stockholm Branch, PO Box 240, S-171 77 Stockholm, Sweden

Departments of † Cell and Molecular Biology and ‡ Neuroscience, Karolinska Institute, S-171 77 Stockholm, Sweden

§ Ligand Pharmaceuticals, 10275 Science Center Drive, San Diego, California 92121, USA

Retinoids regulate gene expression through the action of retinoic acid receptors (RARs) and retinoid-X receptors (RXRs), which both belong to the family of nuclear hormone receptors^{1,2}. Retinoids are of fundamental importance during development³, but it has been difficult to assess the distribution of ligand-activated receptors *in vivo*. This is particularly the case for RXR, which is a critical unliganded auxiliary protein for several nuclear receptors, including RAR¹, but its ligand-activated role *in vivo* remains uncertain. Here we describe an assay in transgenic mice, based on the expression of an effector fusion protein linking the ligand-binding domain of either RXR or RAR to the yeast Gal4 DNA-binding domain, and the *in situ* detection of ligand-activated effector proteins by using an inducible transgenic *lacZ* reporter gene. We detect receptor activation in the spinal cord in a pattern that indicates that the receptor functions in the maturation of limb-innervating motor neurons. Our results reveal a specific activation pattern of Gal4-RXR which indicates that RXR is a critical *bona fide* receptor in the developing spinal cord.

Ligands for retinoid receptors are all-*trans* retinoic acid (RA), which binds to RAR, and 9-*cis* RA, which binds both RAR and

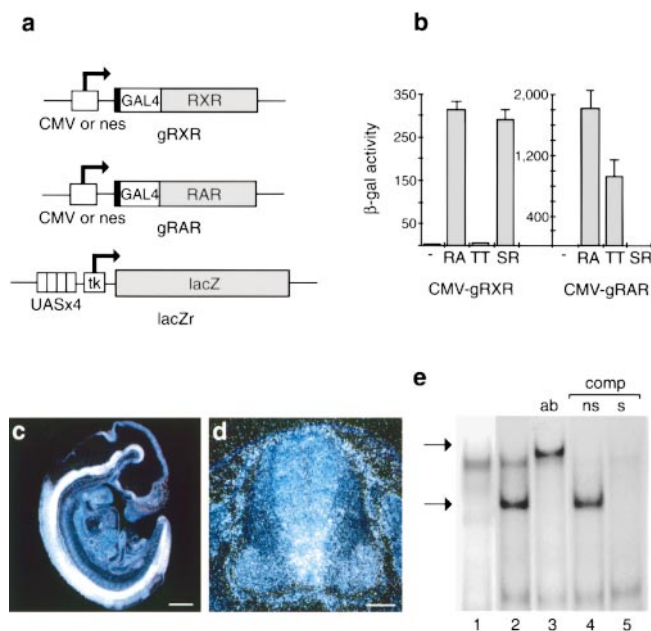


Figure 1 Transgenic effector-reporter assay. **a**, Effector constructs (gRXR and gRAR) contain the ligand-binding domain of RAR and RXR, respectively, fused to a viral HA epitope (black bar) followed by the DNA-binding domain of the yeast Gal4 transcription factor. Cytomegalovirus (CMV) and rat nestin (nes) expression vectors were used for transfection and for transgenic expression, respectively. The reporter transgene (*lacZr*) contains a bacterial β -galactosidase (*lacZ*) gene linked to a minimal thymidine kinase promoter of herpes simplex virus (tk), and Gal4-specific binding sites (UAS $\times$ 4). **b**, β -Galactosidase activity is detected when CMV-gRXR and *lacZr* are co-transfected in the presence of all-trans RA (RA; 1 μ M) or the RXR-selective ligand SR11237 (SR; 1 μ M), but not in presence of the RAR-selective ligand TTNPB (TT; 0.1 μ M) or in the absence of ligand (-). In contrast, gRAR is activated by all-trans RA and TTNPB but not by SR11237. Bars represent mean of four determinations $\pm$ s.d. **c, d**, *In situ* hybridization analysis of gRXR mRNA distribution in E11.5 gRXR-positive embryos. Sagittal (c) and transverse (d) sections of the spinal cord are shown. **e**, Gel-shift analysis of extracts from E10.5 wild-type (lane 1) or gRXR-transgenic (lane 2; lower arrow) embryos incubated with a 32 P-labelled UAS probe. Formation of the protein-DNA complex is inhibited in the presence of excess (50-fold) unlabelled specific UAS competitor ('comp') DNA (s; lane 5), but not by nonspecific DNA (ns; lane 4). The complex is supershifted by anti-HA antiserum (lane 3; ab, top arrow). Scale bars in c, d are 1 mm and 100 μ m, respectively.

RXR². RAR needs to heterodimerize with RXR in order to bind DNA efficiently both *in vivo* and *in vitro*; these heterodimers play key roles *in vivo*³⁻⁵. High levels of all-trans RA have been detected in the embryo, particularly in the spinal cord^{6,7}. In contrast, the function of ligand-activated RXR has remained elusive as 9-cis RA is harder to detect⁷, and because RXR acts as an unliganded heterodimerization partner for several nuclear receptors¹. Nevertheless, the simultaneous ligand activation of RAR and RXR synergistically promotes several physiological responses *in vitro*⁸⁻¹⁴, so to determine whether endogenous RXR-activating ligands exist *in vivo* and to understand the spatiotemporal relation between the regulation of RAR and RXR signalling, we have used transgenic mice for investigating these issues in the developing central nervous system (CNS).

The Gal4 fusion proteins gRXR and gRAR (Fig. 1a) can activate a reporter transgene construct (*lacZr*; Fig. 1a) that contains Gal4 DNA-binding sites upstream of a promoter and a bacterial *lacZ* gene. Transfection of gRXR and gRAR in expression vectors (CMV-gRXR and CMV-gRAR) efficiently induced *lacZ* expression in response to high doses of all-trans RA (Fig. 1b), which results in activation of both receptors owing to partial isomerization of all-trans RA into 9-cis RA¹⁵. In contrast, RAR- and RXR-selective synthetic retinoids specifically activate gRAR and gRXR, respectively.

To detect ligands *in vivo*, we expressed gRXR and gRAR effector proteins in transgenic mice. To analyse the function of retinoid signalling during development of the CNS, gRAR and gRXR were cloned into a rat nestin promoter and enhancer expression vector¹⁶, generating nes-gRXR and nes-gRAR, respectively. *In situ* hybridization with a Gal4-specific probe detects high levels of gRXR messenger RNA in the developing CNS at embryonic day 11.5 (E11.5; Fig. 1c, d). The DNA-binding capacity of gRXR protein was analysed in a gel retardation assay using a 32 P-labelled Gal4 DNA-binding site as probe (Fig. 1e). A shifted complex was formed only in gRXR-positive whole-embryo protein extracts (Fig. 1e: compare lanes 1 and 2, lower arrow).

Transgenic embryos carrying effector and reporter genes either individually or in combination were examined for expression of β -galactosidase by whole-mount X-gal histochemistry (Fig. 2). Our analysis focused on stages from E10.5 onwards, when the nestin promoter is highly active. β -Galactosidase expression was induced in gRAR/*lacZr* and gRXR/*lacZr* embryos but not in embryos carrying *lacZr* alone (Fig. 2g). X-gal staining was detected in specific regions of the spinal cord of gRXR/*lacZr* embryos. At E10.5,

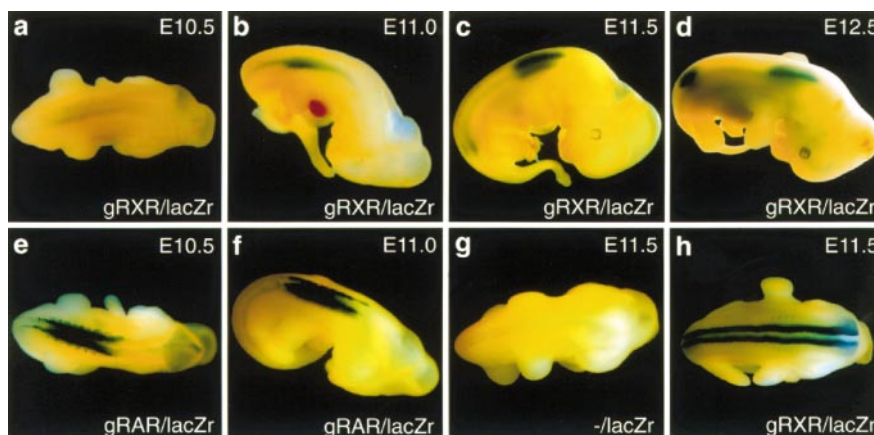


Figure 2 Induction of β -galactosidase expression in transgenic mice. β -Galactosidase expression is specifically induced in gRXR/*lacZr* (a-d) and gRAR/*lacZr* (e, f) embryos at E10.5 (a, e), E11.0 (b, f), E11.5 (c) and E12.5 (d). In *lacZr* embryos (g), no staining is detected. **a-d**, At E10.5, a faint signal is seen at the forelimb level of the spinal cord of gRXR/*lacZr* embryos (a). The intensity

increases at E11.0 (b), and at E11.5 a signal appears at the hindlimb level (c). At E12.5, staining in the forelimb region has declined, but remains high at hindlimb level (d). **e, f**, Spinal cord staining appears already at E10.5 (e) and stays high at E11.0 (f) in gRAR/*lacZr*-positive embryos. **h**, Expanded reporter gene induction along the spinal cord at 12 hours after fetal treatment with all-trans RA.

gRXR-dependent reporter induction was barely detectable, whereas at E11.0 staining was evident at the forelimb level of the spinal cord (Fig. 2a, b). Staining intensified at E11.5, when a weak signal was also detected at hindlimb level of the spinal cord (Fig. 2c). At subsequent stages, β -galactosidase induction intensified at the hindlimb level and peaked at E12.5 (Fig. 2d, and data not shown). This pattern is consistent with the later development of hindlimbs relative to forelimbs. A distinct but related pattern was detected in gRAR/lacZr embryos. Reporter activity significantly preceded that induced by gRXR (Fig. 2e, f). Thus, gRAR-dependent induction was already detected at E10.5 at the forelimb level of the spinal cord (Fig. 2e), whereas gRXR-induced reporter expression appeared at E11.0 (Fig. 2b). To verify that the reporter gene could be induced at other levels of the spinal cord as well, transgenic embryos at E11 were treated *in utero* with high doses of all-*trans* RA and analysed 12 hours post-treatment (Fig. 2h). In gRXR/lacZr embryos, strong β -galactosidase staining was evenly distributed throughout the spinal cord; however, staining was unaccountably weak in the developing brain, despite expression of the effector gene in this region (Fig. 2h, and data not shown).

Our observation that β -galactosidase expression is concentrated at the cervical and lumbar segments of the spinal cord is consistent with previous results, including transgene *lacZ* reporter gene analyses, indicating that these regions are 'hot spots' for retinoid

synthesis^{17–22}. Furthermore, a retinaldehyde dehydrogenase (RALDH-2), considered to be critical for RA synthesis, is expressed in the embryonic ventral spinal cord specifically at these levels^{23–26}; however, these studies did not distinguish between RAR- and RXR-activating ligands, whereas our effector-reporter strategy was designed to assess specific activation by RXR and RAR.

In cervical sections from X-gal-stained transgenic embryos, the patterns of β -galactosidase expression revealed a ventral-to-dorsal spatiotemporal shift. In gRAR/lacZr-positive embryos, staining was most intense in the ventral spinal cord at E10.5 (Fig. 3a) and was later detected in the dorsal region (Fig. 3b). A similar, but delayed, shift was seen in gRXR/lacZr embryos in which lacZr was highly induced ventrally at E11.0 (Fig. 3d, e), and staining increased in dorsal regions at later stages (E11.5, E12.5; Fig. 3f, and data not shown). Taken together, our results show that gRAR and gRXR are both active in spatially similar but temporally unique patterns in the ventral and dorsal spinal cord.

Comparison of RXR and RAR activation patterns may reveal the nature of the natural RXR ligand. The spatial correlation of gRXR- and gRAR-dependent activation indicate that both RAR and RXR ligands are generated in a common biosynthetic pathway. Presumably, these retinoids are identical to all-*trans*- and 9-*cis* RA, respectively.

At stages when both gRAR and gRXR induce β -galactosidase

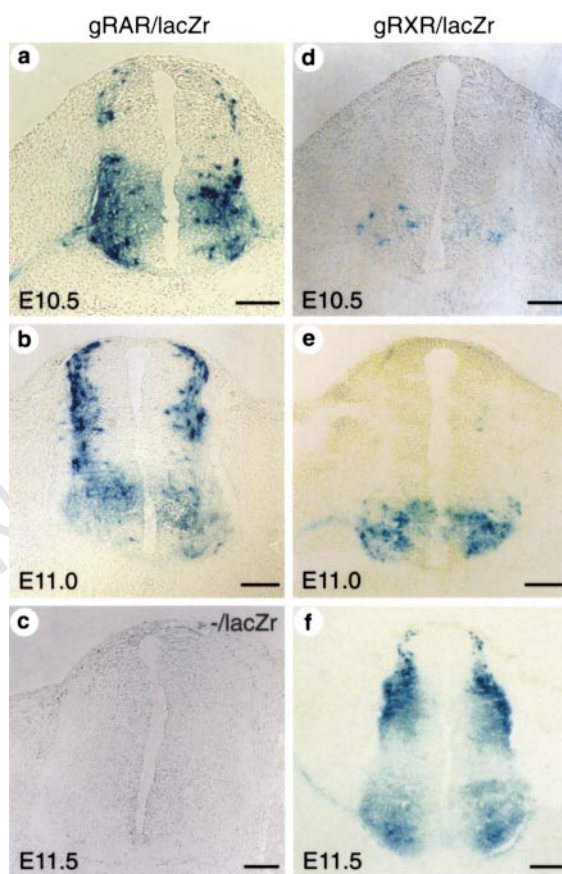


Figure 3 Spatiotemporal shift of β -galactosidase-positive cells in the developing spinal cord. Trans-sections of X-gal-stained gRAR/lacZr (a, b) and gRXR/lacZr (d–f) embryos reveal a ventral-to-dorsal spatiotemporal shift of β -galactosidase-positive cells. a, b, Strongly blue-stained cells of gRAR/lacZr-embryos are detected ventrally in the E10.5 spinal cord (a). At E11.0 staining is seen mainly in dorsal spinal cord (b). c, No blue cells are detected in the -/lacZr embryo. d–f, Few positive cells are present in the ventral E10.5 spinal cord of gRXR/lacZr embryos (d). An increased number of strongly positive cells is observed at E11.0 (e). At E11.5, staining is mainly observed dorsally (f). Scale bars, 100 μ m.

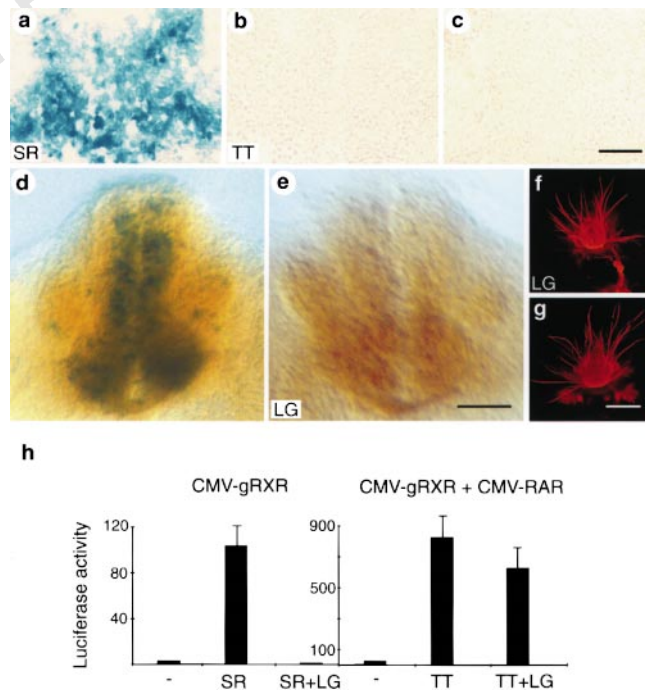


Figure 4 Induction and inhibition of β -galactosidase expression in E10.5 spinal cord explant cultures of gRXR/lacZr transgenic embryos. Explants derived from thoracic (a–c) and cervical (d–g) segments were cultured in presence of SR11237 (1 μ M; $n = 4$) (a), TTNPB (0.1 μ M; $n = 4$) (b), vehicle ($n = 9$) (c, d), or LG100849 (2 μ M; $n = 5$) (e). a–c, β -Galactosidase-positive cells, normally absent from thoracic spinal cord (c), were detected in SR11237-treated (a) but not in TTNPB-treated (b) cultures. d, e, The endogenous β -galactosidase signal, appearing at the forelimb level of the spinal cord (d), was blocked by LG100849 (e). f, g, Explants cultured for 5 d in the presence of antagonist (f) or vehicle (g) were analysed for neurite outgrowth by immunodetection of the neural marker Tuj-1. h, gRXR activation is blocked by LG100849, whereas there is only a minor reduction of activation when RAR is overexpressed with gRXR. The indicated effector constructs (CMV-gRXR or CMV-gRXR+CMV-RAR) and Gal4-responsive luciferase reporter plasmid were co-transfected with vehicle (–), SR11237 (SR; 0.1 μ M) or TTNPB (TT; 0.01 μ M) in the absence or presence of LG100849 (LG; 1 μ M). Bars represent mean of 4 determinations $\pm$ s.d. Scale bars in c, e, g, 100 μ m.

expression (E11.0; Fig. 2b, f), the dorsoventral distribution of positive cells is distinct, with gRAR signalling being most prominent dorsally and that of gRXR ventrally (Fig. 3), supporting the selectivity in activation of gRXR and gRAR. Therefore, it is unlikely that reporter induction by gRXR is a result of heterodimerization with endogenous RAR. However, the specificity of reporter gene induction in gRXR/lacZr transgenic mice was confirmed in several ways: first, transgenic spinal cord explants from thoracic segments of E10.5 embryos, which do not contain X-gal positive cells in gRXR/lacZr transgenic embryos, were treated with RXR- or RAR-selective retinoids, known as SR11237 and TTNPB respectively (Fig. 1b). After 24 hours, only explants cultured in the presence of SR11237 showed strong X-gal staining (Fig. 4a–c). Second, gRXR/lacZr transgenic explants were treated for 24 hours with a selective RXR antagonist, LG100849 (see Methods). In controls, these explants expressed high levels of β -galactosidase (Fig. 4d), whereas the antagonist completely blocked the appearance of blue cells (Fig. 4e). Cellular viability in spinal cord explants cultured in the presence of antagonist was controlled by examining neurite outgrowth and testing for the presence of Islet 1 immunoreactivity (Fig. 4f, g, and data not shown). Furthermore, the antagonist did not abolish activation by RXR heterodimers in transfected cells expressing Gal4–RXR together with RAR (Fig. 4h). Finally, gRXR/lacZr transgenic embryos treated *in utero* with teratogenic doses of the RAR-selective agonist TTNPB did not show increased β -galactosidase expression (data not shown). Thus, although we cannot exclude the possibility that a putative heterodimerization partner and its ligand are promoting activation through gRXR, our results indicate that the pattern of X-gal-stained cells in the spinal cord of gRXR embryos is generated by an endogenous ligand activating RXR *in vivo*.

We have provided evidence for RXR signalling *in vivo* that supports a role for RXR-activating retinoids in motor innervation of the limbs. A recent study does indeed implicate RAR and RXR signalling in the development of a population of limb-innervating motor neurons²² and new genetic evidence further supports a role for RXR signalling *in vivo*³¹. The transgenic approach described here is a general strategy that should prove useful in defining activation patterns of other nuclear receptors, including those whose ligands are still unidentified. □

Methods

Plasmid construction. The transgene reporter plasmid lacZr contains four repeats of Gal4-specific binding sites (upstream activating sequences, UAS), followed by a minimal thymidine kinase promoter linked to the lacZ gene. The effector constructs contain the yeast Gal4 DNA-binding domain (amino acids 1–147), flanked by a 9-amino-acid-long haemagglutinin epitope of influenza virus (HA) at the N terminus, and the ligand-binding domain of the human RAR α or RXR α at the C terminus. The resulting coding sequences were cloned into pCMX or pNesSV40pA expression vectors^{27,28} (details available from the authors).

Transfections. Transfection of human JEG-3 choriocarcinoma cells was done as described²⁷. Ligands added were: all-*trans* RA (1 μ M), TTNPB ((E)-4-[2-(5,5,8,8-tetramethyl-5,6,7,8 tetrahydro-2-naphthalenyl)-1-propenyl] benzoic acid; 0.1 μ M) or SR11237 (1 μ M). For antagonist experiments, 1 μ M LG100849 was included with SR11237 (0.1 μ M) or TTNPB (0.01 μ M). LG100849 is a selective RXR antagonist and binds with high affinity to RXR α , RXR β and RXR γ , but not significantly to any of the RARs at concentrations below 10 μ M (I. G. Schulman, R.P.B. and R.A.H., manuscript in preparation).

Transgenic animals. Transgenic mice were produced as described²⁸. Founder lines were established for nes-gRXR and lacZr, whereas nes-gRAR embryos of different stages were obtained by transient injection as the nes-gRAR transgene was not transmitted through the germ line. Separate nes-gRXR effector founder lines showed reproducible transgenic expression, as judged from *in situ* hybridization (three out of four lines), and gave virtually identical X-gal staining patterns when crossed with reporter mice (two lines). For prenatal

retinoid treatment, pregnant mice were gavage-fed 11 days post-coitum with all-*trans* RA dissolved in 300 μ l corn oil at a concentration of 20 mg kg⁻¹ body weight.

Explant cultures. Explants were cultured for 1–5 days as described²⁹ in Dulbecco's modified Eagle's medium containing 0.2% ethanol, 1 μ M SR11237, 0.1 μ M TTNPB or 2 μ M LG100849. Cultures were fixed and X-gal-stained or analysed by immunohistochemistry.

Gel-shift analysis. Total embryonic protein extracts were prepared from E10.5 wild-type or gRXR transgenic embryos. Gel-shift analyses were done as described, using a ³²P-labelled UAS oligonucleotide probe (5'-AGC TCG ACG GAG TAC TGT CCT CCG TCG A-3') alone or in the presence of 50-fold excess unlabelled UAS- or nonspecific double-stranded DNA oligonucleotide, or monoclonal anti-HA antibodies.

Histochemical analysis. Embryos were analysed at different stages from E10.5 to E13.5, with the morning of plugging being counted as 0.5 days post-coitum (E0.5); stages were confirmed on the basis of morphology. X-gal (5-bromo-4-chloro-3-indolyl- β -D-galactopyranoside) staining²⁸ induced by gRAR and gRXR was equal, as judged by staining intensity. For indirect immunohistochemistry and *in situ* hybridization, explants and embryos were fixed in 2–4% paraformaldehyde. Primary monoclonal antibodies were incubated overnight at the following dilutions: Tuj-1 (anti- γ -tubulin), 1:250; 4D5 (anti-Islet 1/2), 1:50. For *in situ* hybridization, sections were incubated overnight with ³⁵S-labelled oligonucleotide probes and processed as described³⁰. The Gal4-specific probe had the sequence 5'-AAA ACC AAA AGG TCT CCG CTG ACT AGG GCA CAT CTG ACA GAA GTG GAA-3'.

Received 9 March; accepted 7 July 1998.

- Mangelsdorf, D. J. & Evans, R. M. The RXR heterodimers and orphan receptors. *Cell* **83**, 841–850 (1995).
- Kastner, P., Mark, M. & Chambon, P. Nonsteroid nuclear receptors: what are genetic studies telling us about their role in real life? *Cell* **83**, 859–869 (1995).
- Marshall, H. *et al.* A conserved retinoic acid response element required for early expression of the homeobox gene Hoxb-1. *Nature* **370**, 567–571 (1994).
- Kastner, P. *et al.* Genetic evidence that the retinoid signal is transduced by heterodimeric RXR/RAR functional units during mouse development. *Development* **124**, 313–326 (1997).
- Dupé, V. *et al.* *In vivo* functional analysis of the Hoxa-1 3' retinoic acid response element (3' RARE). *Development* **124**, 399–410 (1997).
- Wagner, M., Thaller, C., Jessell, T. M. & Eichele, G. Polarizing activity and retinoid synthesis in the floorplate of the neural tube. *Nature* **345**, 819–822 (1990).
- Horton, C. & Maden, M. Endogenous distribution of retinoids during normal development and teratogenesis in the mouse embryo. *Dev. Dynam.* **202**, 312–323 (1995).
- Durand, B. *et al.* Activation function 2 (AF2) of retinoic acid receptor and 9-*cis* retinoic acid receptor: presence of a conserved autonomous constitutive activation domain and influence of the nature of the response element on AF-2 activity. *EMBO J.* **13**, 5370–5382 (1994).
- Yang, Y., Minucci, S., Ozato, K., Heyman, R. A. & Ashwell, J. D. Efficient inhibition of activation-induced Fas ligand up-regulation and T cell apoptosis by retinoids requires occupancy of both retinoid X receptors and retinoic acid receptors. *J. Biol. Chem.* **270**, 18672–18677 (1995).
- Roy, B., Taneja, R. & Chambon, P. Synergistic activation of retinoic acid (RA)-responsive genes and induction of embryonic carcinoma cell differentiation by an RA receptor α (RAR α), RAR β , or RAR γ -selective ligand in combination with a retinoid X receptor-specific ligand. *Mol. Cell. Biol.* **15**, 6481–6487 (1995).
- Bissonnette, R. P. *et al.* 9-*cis* retinoic acid inhibition of activation-induced apoptosis is mediated via regulation of fas ligand and requires retinoic acid receptor and retinoid X receptor activation. *Mol. Cell. Biol.* **15**, 5576–5585 (1995).
- Botling, J., Castro, D. S., Ösberg, E., Nilsson, K. & Perlmann, T. RAR/RXR heterodimers can be activated through both subunits providing a basis for synergistic transactivation and cellular differentiation. *J. Biol. Chem.* **272**, 9443–9449 (1997).
- Minucci, S. *et al.* Retinoid X receptor (RXR) within the RXR–retinoic acid receptor heterodimer binds its ligand and enhances retinoid-dependent gene expression. *Mol. Cell. Biol.* **17**, 644–655 (1997).
- Lu, H.-C., Eichele, G. & Thaller, C. Ligand-bound RXR can mediate retinoid signal transduction during embryogenesis. *Development* **124**, 195–203 (1997).
- Allenby, G. *et al.* Retinoic acid receptors and retinoid X receptors: interactions with endogenous retinoic acids. *Proc. Natl Acad. Sci. USA* **90**, 30–34 (1993).
- Zimmerman, L. *et al.* Independent regulatory elements in the nestin gene direct transgene expression to neural stem cells or muscle precursors. *Neuron* **12**, 11–24 (1994).
- Rossant, J., Ziringib, R., Cado, D., Shago, M. & Giguère, V. Expression of a retinoic acid response element–*hsp*lacZ transgene defines specific domains of transcriptional activity during mouse embryogenesis. *Genes Dev.* **5**, 1333–1344 (1991).
- Reynolds, K., Mezey, E. & Zimmer, A. Activity of the β -retinoic acid receptor promoter in transgenic mice. *Mech. Dev.* **36**, 15–29 (1991).
- Mendelsohn, C., Ruberte, E., LeMeur, M., Morris-Kay, G. & Chambon, P. Developmental analysis of the retinoic acid-inducible RAR- β promoter in transgenic animals. *Development* **113**, 723–734 (1991).
- Wagner, M. & Jessell, T. M. Regional differences in retinoid release from embryonic neural tissue detected by an *in vitro* reporter assay. *Development* **116**, 55–66 (1992).
- Colbert, M. C., Linney, E. & LaMantia, A.-S. Local sources of retinoic acid coincide with retinoid-mediated transgene activity during embryonic development. *Proc. Natl Acad. Sci. USA* **90**, 6572–6576 (1993).
- Sockanathan, S. & Jessell, T. M. Motor neuron-derived retinoid signaling specifies the subtype identity of spinal motor neurons. *Cell* **94**, 503–514 (1998).
- McCaffery, P. & Dräger, U. C. Hot spots of retinoic acid synthesis in the developing spinal cord. *Proc. Natl Acad. Sci. USA* **91**, 7194–7197 (1994).

24. Wang, X., Penzes, P. & Napoli, J. L. Cloning of a cDNA encoding an aldehyde dehydrogenase and its expression in *Escherichia coli*. Recognition of retinal as substrate. *J. Biol. Chem.* **271**, 16288–16293 (1996).
25. Zhao, D. *et al.* Molecular identification of a major retinoic-acid-synthesizing enzyme, a retinaldehyde-specific dehydrogenase. *Eur. J. Biochem.* **240**, 15–22 (1996).
26. Niederreither, K., McCaffery, P., Dräger, U. C., Chambon, P. & Dollé, P. Restricted expression and retinoic acid-induced downregulation of the retinaldehyde dehydrogenase type 2 (RALDH-2) gene during mouse development. *Mech. Dev.* **62**, 67–78 (1997).
27. Perlmann, T. & Jansson, L. A novel pathway for vitamin A signaling mediated by RXR heterodimerization with NGFI-B and Nurr1. *Genes Dev.* **9**, 769–782 (1995).
28. Nilsson, E. & Lendahl, U. Transient expression of a human β -actin promoter/*lacZ* gene introduced into mouse embryos correlates with a low degree of methylation. *Mol. Reprod. Dev.* **34**, 149–157 (1993).
29. Lumsden, A. G. S. & Davies, A. M. Earliest sensory nerve fibres are guided to peripheral targets by attractants other than nerve growth factor. *Nature* **306**, 786–788 (1983).
30. Zetterström, R., Solomin, L., Mitsiadis, T., Olson, L. & Perlmann, T. RXR heterodimerization and developmental expression distinguish the orphan nuclear receptors NGFI-B, Nurr1 and Nor1. *Mol. Endocrinol.* **10**, 1656–1666 (1996).
31. Mascréz, B. *et al.* The RXR α ligand-dependent activation function 2 (AF-2) is important for mouse development. *Development* (in the press).

Acknowledgements. We thank S. Sockanathan, T. Jessell and P. Chambon for communicating unpublished results and for discussion; T. Jessell and P. Ljungdahl for Islet-1 and HA antibodies; A. Mata for transfections; E. Nilsson, K. Jansson, Karin Lundströmer and E. Lindqvist for technical support; L. Foley for SR11237; and R. Pettersson for comments. This work was supported by the Swedish Medical Research Council, and the US Public Health Service.

Correspondence and requests for materials should be addressed to T.P. (e-mail: thomas.perlmann@licr.ki.se).

Position and orientation of the globular domain of linker histone H5 on the nucleosome

Y.-B. Zhou*, Sue Ellen Gerchman†, V. Ramakrishnan‡, Andrew Travers§ & Serge Muyldermans*

* Vlaams Interuniversitair Instituut voor Biotechnologie, Vrije Universiteit Brussel, Paardenstraat 65, B-1640 Sint Genesius Rode, Belgium

† Biology Department, Brookhaven National Laboratory, Upton, New York 11973, USA

‡ Department of Biochemistry, University of Utah, School of Medicine, Salt Lake City, Utah 84132, USA

§ MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

It is essential to identify the exact location of the linker histone within nucleosomes, the fundamental packing units of chromatin, in order to understand how condensed, transcriptionally inactive chromatin forms. Here, using a site-specific protein–DNA photocrosslinking method¹, we map the binding site and the orientation of the globular domain of linker histone H5 on mixed-sequence chicken nucleosomes. We show, in contrast to an earlier model², that the globular domain forms a bridge between one terminus of chromatosomal DNA and the DNA in the vicinity of the dyad axis of symmetry of the core particle. Helix III of the globular domain binds in the major groove of the first helical turn of the chromatosomal DNA, whereas the secondary DNA-binding site on the opposite face of the globular domain of histone H5 makes contact with the nucleosomal DNA close to its midpoint. We also infer that helix I and helix II of the globular domain of histone H5 probably face, respectively, the solvent and the nucleosome. This location places the basic carboxy-terminal region of the globular domain in a position from which it could simultaneously bind the nucleosome-linking DNA strands that exit and enter the nucleosome.

The linker histone H1 and its avian erythrocyte variant H5 facilitate the folding of the higher-order 300-Å chromatin fibre³ whose fundamental unit is the chromatosome which contains all of the core histones, one molecule of linker histone and 168 basepairs (bp) of DNA⁴. The linker histones contain basic amino- and C-terminal tails flanking a central globular domain⁵, which consists of a three-helix bundle (helices I–III) with a C-terminal β -hairpin^{6,7}. Because of its structural similarity to the helix–turn–helix proteins⁸, helix III of GH5, the globular domain of H5, was

proposed to bind in the major groove of DNA, with the β -hairpin lying over an adjacent minor groove^{6,8}. The opposite face of GH5 contains a less-defined secondary DNA-binding site formed by a clustering of highly conserved basic residues⁶. Both DNA-binding sites are necessary for chromatosome formation and for binding two duplex DNA molecules⁹.

The exact binding site of GH5 on the nucleosome remains controversial¹⁰. Initial models indicated that the globular domain bound to the central gyre at the nucleosome pseudodyad^{4,5}. However, both neutron diffraction of chromatosomes and the identification of a conserved sequence close to one end of chromatosomal DNA^{11–13} support an asymmetric position for GH5, spanning the DNA gyres between one terminus and the midpoint. An alternative asymmetric model, derived from studies of a chromatosome reconstituted on *Xenopus borealis* somatic 5S ribosomal DNA^{2,14}, proposed that GH5 makes a single contact with DNA on the inside of the DNA gyre at an internal position that is 65 bp from the dyad.

To determine the location of the globular domain of H5 in bulk chromatin, we prepared ‘random-sequence’ chromatosomes from chicken erythrocytes. The endogenous linker histones were first removed selectively with resin¹⁵, and the chromatosomal DNA was end-labelled¹⁶. DNase I digestions of these H1/H5-depleted chromatosomes showed strong cutting sites around nucleotides at positions 21, 31, 52, 63, 105, 115, 135 and 146 from the 5′ extremity (Fig. 1). This pattern of bands is shifted by 10 bp relative to the characteristic ladder obtained from similar experiments on nucleosome core particles using DNase I digestions¹⁶, and it shows that the H1/H5 depletion and the end-labelling conditions did not result in the histone octamer sliding towards a terminal position. Indeed, the histone octamer on the 168-bp DNA retains the symmetrical cleavage pattern found in native chromatosomes^{4,17}.

We then reconstituted these H1/H5-depleted chromatosomes with either recombinant GH5 or each of six GH5 mutants containing

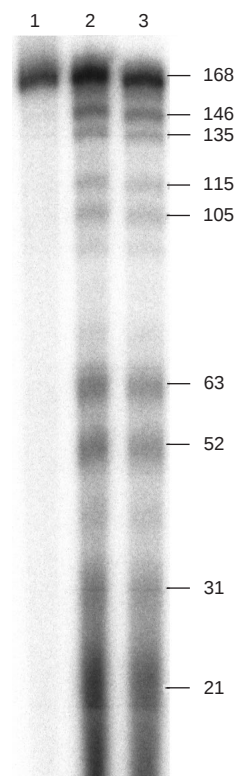


Figure 1 DNase I digestion of H1/H5-depleted chromatosomes. The linker-histone-depleted chromatosomes prepared from chicken erythrocytes were end-labelled with ³²P (lane 1) and further digested with DNase I for 10 or 20 s (lanes 2 and 3, respectively), before loading on a 6% denaturing gel. The sizes (in nucleotides) of the most intense bands are marked on the right.

a single Ser-to-Cys substitution. Each of these substitutions was located on a different face of the GH5 domain: mutant S29C immediately precedes helix I; mutant S41C is within the secondary DNA-binding site; mutant S45C is part of the β -sheet; mutants S49C and S56C are at the N- and C-terminal ends of helix II, respectively; and mutant S71C is within the putative major groove binding helix III⁸.

All of the mutant proteins retained their capacity to bind to depleted chromatosomes (Fig. 2a). In the absence of dithiothreitol (DTT), both the S41C and S56C proteins dimerized strongly; S71C failed to form intermolecular disulphide bonds; and all other mutants dimerized to an intermediate extent (Fig. 2b). A nucleoprotein gel indicated that only S41C and S45C could bind to a chromatosome as dimers (Fig. 2a, lanes without DTT). This result shows that residues 41 and 45 of GH5 should be accessible to the solvent when bound in a chromatosome.

To identify GH5 residues that are close to the nucleosomal DNA, each single Cys residue was first reacted with *p*-azidophenacylbromide¹. On photoactivation of the azido group these conjugates can then crosslink to DNA in van der Waals contact. We observed that only two of the resulting mutant-azidophenacyl

conjugates, S49C-azidophenacyl and S56C-azidophenacyl, lost their ability to interact with the depleted chromatosomes (Fig. 2c). Failure of these two GH5 conjugates to form chromatosomes might result from steric hindrance at the nucleosome-binding site, for example when S49 and S56 would face the nucleosome particle. Unfolding of GH5 on modification is another possibility, although we consider this less likely because these conjugates were still able to bind specifically to a fourway DNA junction (Y.-B.Z., manuscript in preparation). From these results and those of the dimerization studies, we conclude that helix II of the globular domain faces the interior of the chromatosome, and therefore helix I is on the outside. This orientation is consistent with the protection of K52 and K55 in helix II from reductive methylation in chromatin containing linker histone¹⁸.

We determined the location of the azidophenacyl group within the reconstituted chromatosomes for all of the azidophenacyl-conjugated mutants that retained chromatosome binding. On exposure to ultraviolet light, S45C-azidophenacyl was not cross-linked to the chromatosomal DNA, indicating that, consistent with the previous conclusion, this residue is probably directed towards the solvent and away from the chromatosome. In contrast, S41C,

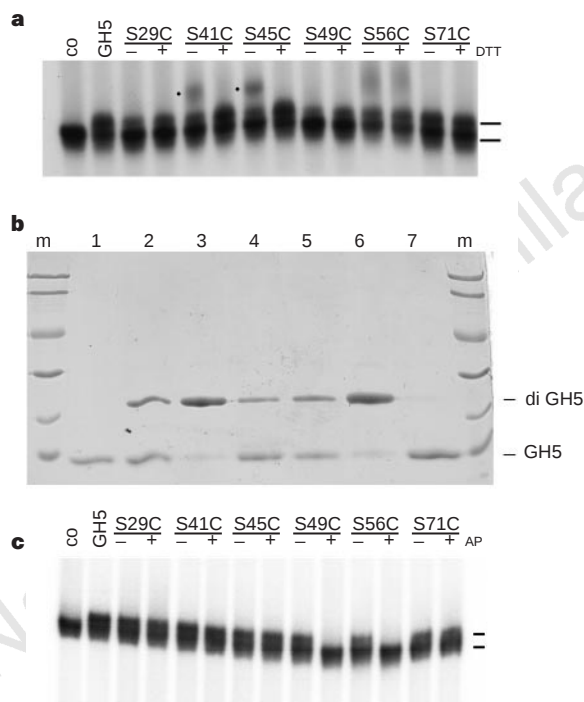


Figure 2 Characterization of the complex formation between GH5 mutants and linker-histone-depleted chromatosomes and of dimer formation by the mutant proteins. **a**, The end-labelled chromatosomes (5 pmol) were run as nucleoprotein sample (co) or as GH5 chromatosome reconstituted with native recombinant GH5 (GH5). The different GH5 mutants (S29C, S41C, S45C, S49C, S56C or S71C), which were added at a 1:1 stoichiometry to the chromatosomes, were incubated without (-) or with (+) 1 mM DTT before separation through gel electrophoresis. The lower line on the right of the figure indicates the position of the linker-histone-depleted chromatosome; the upper line denotes the position to which the GH5-reconstituted chromatosomes migrate. The signal at the position of the black dot corresponds to the chromatosomes with the GH5 dimer. (Particles that contain two nucleosomes migrate much more slowly in these 8% polyacrylamide gels, and are outside the region shown.) **b**, SDS-protein gel of the native recombinant GH5 (lane 1), and the S29C (lane 2), S41C (lane 3), S45C (lane 4), S49C (lane 5), S56C (lane 6) and S71C (lane 7) GH5 mutants. Reducing agents were omitted from the sample solution. The lanes labelled 'm' are molecular-size markers (M_r) of 90, 67, 45, 30, 20 and 14K. **c**, Electrophoresis of the chromatosomes with the GH5 mutants before (-) or after (+) conjugation to the photoactivatable AP crosslinker. Notations are as in (a).

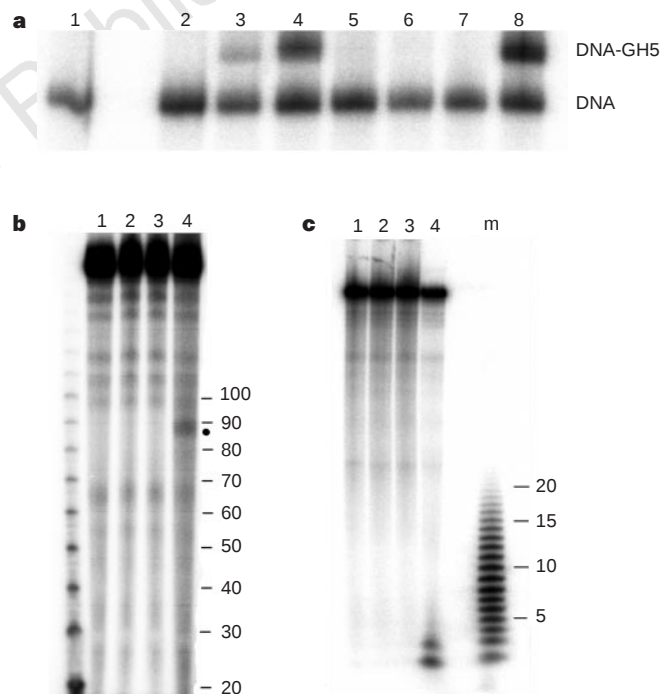


Figure 3 Analysis of GH5-azidophenacyl crosslinked to chromatosomal DNA. **a**, The GH5-azidophenacyl conjugates added to the linker-histone-depleted chromatosomes were exposed to ultraviolet light and the covalently crosslinked protein-DNA was enriched by phenol extraction and ethanol precipitation before separation on a 6% polyacrylamide gel run in 1 × TBE, 0.1% SDS. The samples obtained without GH5 addition (lane 1) or by adding the native recombination GH5 (lane 2) are used as negative controls. Other lanes contained the azidophenacyl-conjugated materials (lane 3, S29C; lane 4, S41C; lane 5, S45C; lane 6, S49C; lane 7, S56C; and lane 8, S71C). **b**, Denaturing 8% polyacrylamide gel electrophoresis of end-labelled linker-histone-depleted chromatosomes reconstituted with native recombinant GH5 (lane 1) or with S41C GH5 mutant before (lane 2) or after (lanes 3, 4) conjugation with azidophenacyl. The sample shown in lane 3 was not exposed to ultraviolet light, whereas that of lane 4 was crosslinked by exposure to ultraviolet light. All samples were extracted with phenol, precipitated with ethanol and treated with hot piperidine to cleave the DNA at the protein crosslinking point. The sizes in nucleotides are shown on the right, and the black dot denotes the position of the internal cleavage introduced by the S41C-azidophenacyl conjugate. **c**, Denaturing 20% polyacrylamide gel electrophoresis of the experiment with the S71C mutant. Treatment and order of samples are as in (b).

located within the putative secondary DNA-binding site, and S29C and S71C, located in site I, crosslinked at significant levels to chromosomal DNA (Fig. 3a). To map the interaction sites, the DNA at the point of crosslinking was cleaved by alkali at a position immediately adjacent to the crosslink. Cleavage of the S41C crosslinked material generated one major cleavage product of ~86 nucleotides (Fig. 3b). The cleavage is close to the midpoint of chromosomal DNA and indicates that the secondary DNA-binding site interacts with the DNA in the vicinity of the nucleosome pseudodyad (Fig. 4).

Analysis of the cleavage products of the conjugated S71C–azidophenacyl in a 20% acrylamide gel revealed bands of only a few nucleotides in length (Fig. 3c), indicating that helix III is positioned at one terminus of chromosomal DNA (Fig. 4). We saw no distinct cleavage product from material that was crosslinked

with S29C–azidophenacyl. From the orientation of GH5, we would expect that crosslinking of S29C–azidophenacyl would yield even shorter DNA fragments than S71C–azidophenacyl, or it would cleave close to the 3' end of the complementary strand, with consequent difficulties of recovery and detection. The AGGA signal sequence¹³ is located in a similar position to the S71C–azidophenacyl crosslink, but it is also possible that this sequence constitutes a binding site for a core histone tail that is dislocated on GH5 binding. From the structure of the core particle¹⁹, we surmise that the N tail of H3 or the C tail of H2A would be likely candidates to extend towards these positions, protecting the 168-bp chromosome fragment from degradation by micrococcal nucleases.

The interaction of helix III of GH5 at one terminus of the chromosome and of a core-histone tail at the other terminus could explain a symmetrical 10-bp extension of DNA protection of core nucleosomal DNA on binding the linker histone. However, H5 binding may result in an asymmetric extension of protection of 20 bp on one side and 0 bp on the other²⁰. In principle, our data are compatible with either symmetric or asymmetric extension. In the latter case, however, we have to assume that DNase I cleavage of the stripped chromosomes is asymmetric with respect to the dyad. In either situation, the major cleavage by S41C–azidophenacyl would be close to the pseudodyad. Our conclusions are also consistent with other results, namely the binding of two DNA duplexes by GH5 (ref. 21); the requirement for both DNA-binding sites for chromosome formation⁹; the footprinting of GH5 in the vicinity of the dyad^{22,23}; pyrimidine dimer formation in native chromatin²⁴; neutron-scattering studies¹¹; and the crosslinking of H25 to DNA in chromatin²⁵.

The position and orientation of GH5 on the chromosome places H25 very close to one terminus of the chromosomal DNA and we suggest that this residue defines one of the pause sites for micrococcal nuclease. This places the basic N-terminal tail of intact H5 near linker DNA, although it could also make contact with a neighbouring chromosome. Extension of the C-terminal end of GH5 (Fig. 4c,d) would allow the extended basic C-terminal tail to contact one of both of the entering and the exiting duplexes²⁶.

Our results differ fundamentally from a model proposed for the organization of the chromosome formed on the *X. borealis* somatic 5S gene^{2,14,27}. That model prescribes a location for GH5 that cannot make contact with the dyad because it is asymmetrically located inside the DNA superhelix. While it is possible that this model represents a special characteristic of the 5S nucleosome, it depends on the assumption that this nucleosome adopts a single, dominant translational position. However, accurate mapping of nucleosomal dyads both on this sequence and on the closely related *Xenopus laevis* somatic 5S genes reveals the presence of multiple equivalent translational positions both with and without histone H1 (refs 28, 29). Consequently, the contacts deduced from crosslinking^{2,27} and directed DNA cleavage¹⁴ cannot be assigned without ambiguity to any particular translational position. By contrast, in our experiments the end positions of the DNA are determined by nuclease digestion, which in turn is defined by the structure of the chromosome, so that there is a definite relationship between the ends of the DNA and the position of the dyad. □

Methods

Chromosomes and GH5 preparations. Chromosomes were prepared as described previously¹¹. The linker histones were depleted with resin AG50W-X2¹⁵. The 5' ends of DNA within the chromosomes were end-labelled with polynucleotide kinase (Gibco-BRL) and [γ -³²P]ATP, and the DNase I (Sigma) digestions were done following standard procedures^{4,16}.

The native recombinant GH5 used as controls and the recombinant GH5 mutants were expressed in *Escherichia coli*⁶. The GH5 mutants were reduced by addition of 2 mM DTT, passed over G15 Sephadex in 20 mM Tris pH 7.6, 20 mM KCl and 0.1% Nonidet NP40 or a 0.1 × Sulpho buffer⁷, conjugated to *p*-azidophenacylbromide (Sigma)¹, and gel-filtrated as above to remove the unreacted crosslinking agent. The recombinant GH5 or its azidophenacyl

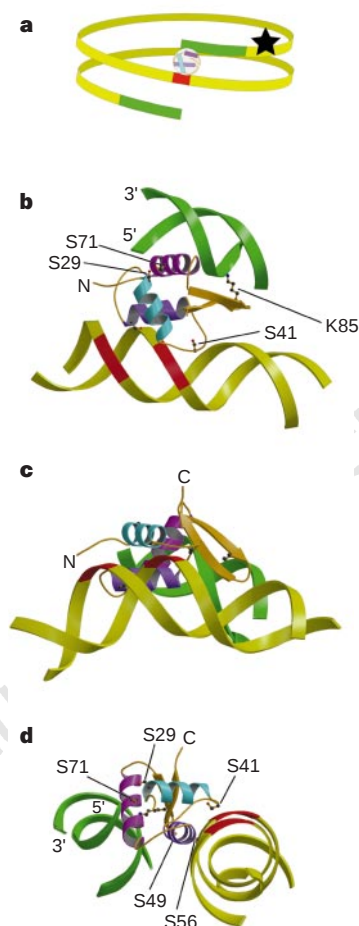


Figure 4 Representation of the GH5 location on a chromosome. The GH5 (circle with helix I, cyan; helix II, purple; helix III, magenta; and turns and β -strands in orange) inserted between the two DNA gyres is positioned and orientated in agreement with our footprinting results. **a**, Representation of the GH5 location on a chromosome. The full two superhelical turns of chromosomal DNA are shown; the black star indicates the position of GH5 as proposed in ref. 2. The 3 bp around the core particle pseudodyad are in red; the 10 bp at the outside of the core-particle DNA are in green; all other DNA is in yellow. **b–d**, A more detailed model of GH5 orientation on the central DNA gyre and the 'entry' chromosomal DNA. The 15 bp around the chromosomal midpoint are shown together with the 'entry' 10 bp (green) of the chromosomal DNA. These figures are schematic and do not take into account possible DNA distortions in the dyad or 'exit' region. The views of **a**, **b** are as seen in the direction of the pseudodyad axis of the core particle; **c** was obtained by rotating **b** through 90° along a horizontal line; the pseudodyad axis runs from top to bottom; **d** shows the rotation of **c** by 90° over its pseudodyad axis.

conjugate was mixed with the end-labelled chromosomes at an estimated stoichiometry of 1:1 in $0.5 \times$ TBE buffer.

Nucleoprotein gels and GH5 footprinting. The nucleoprotein gels contained 8% acrylamide (80:1 acryl:bisacryl) in $0.5 \times$ TBE and up to 20% glycerol. Separation of the linker-histone-depleted chromosomes and the GH5-reconstituted chromosomes was achieved by overnight electrophoresis at 4°C .

The GH5-azidophenacyl conjugates were covalently attached to the chromosomal DNA by exposure of the reconstitutes to a standard ultraviolet transilluminator (312 nm) for 4 min. The crosslinked GH5-DNA complex was extracted with phenol and was precipitated from the phenol phase with ethanol¹. DNA was cleaved at the point of crosslinking by treatment with 1 M piperidine at 90°C for 30 min. The resulting DNA fragments were denatured by addition of formamide and boiling before separation according to size on DNA-sequencing gels containing 7 M urea. The buffer in the lower gel tank of the 20% gel shown in Fig. 3c contained two volumes $1 \times$ TBE and one volume 3 M sodium acetate.

Received 12 March; accepted 13 July 1998.

- Pendegast, P. S., Chen, Y., Ebright, Y. W. & Ebright, R. H. Determination of the orientation of a DNA binding motif in a protein-DNA complex by photocrosslinking. *Proc. Natl Acad. Sci. USA* **89**, 10287–10291 (1992).
- Pruss, D. et al. An asymmetric model for the nucleosome: a binding site for linker histones inside the DNA gyres. *Science* **274**, 614–617 (1996).
- Thoma, F., Koller, T. & Klug, A. Involvement of histone H1 in the organisation of the nucleosome and of the salt dependent superstructure of chromatin. *J. Cell Biol.* **83**, 403–427 (1979).
- Simpson, R. Structure of the chromosome, a chromatin particle containing 160 bp of DNA and all histones. *Biochemistry* **17**, 5524–5531 (1978).
- Allan, J., Crane-Robinson, C. & Aviles, F. X. The structure of histone H1 and its location in chromatin. *Nature* **288**, 675–679 (1980).
- Ramakrishnan, V., Finch, J. T., Graziano, V., Lee, P. L. & Sweet, R. M. Crystal structure of the globular domain of histone H5 and its implications for nucleosome binding. *Nature* **362**, 219–223 (1993).
- Cerf, C. et al. Homo- and heteronuclear two-dimensional NMR studies of the globular domain of histone H1: full assignment, tertiary structure and comparison with the globular domain of histone H5. *Biochemistry* **33**, 11079–11086 (1994).
- Clark, K. L., Halay, E. D., Lai, E. & Burley, S. K. Co-crystal structure of the HNF-3/Fork head DNA-recognition motif resembles histone H5. *Nature* **364**, 412–420 (1993).
- Goytisolo, F. A. et al. Identification of two DNA-binding sites on the globular domain of histone H5. *EMBO J.* **15**, 3421–3429 (1996).
- Crane-Robinson, C. Where is the globular domain of linker histone located on the nucleosome? *Trends Biochem. Sci.* **22**, 75–77 (1997).
- Lambert, S. et al. Neutron scattering studies of chromosomes. *Biochem. Biophys. Res. Commun.* **179**, 810–816 (1991).
- Muyldermans, S. & Travers, A. A. DNA sequence organization in chromosomes. *J. Mol. Biol.* **235**, 855–870 (1994).
- Travers, A. A. & Muyldermans, S. A. DNA sequence for positioning chromosomes. *J. Mol. Biol.* **257**, 486–491 (1996).
- Hayes, J. J. Site-directed cleavage of DNA by a linker histone-Fe(II) EDTA conjugate: localization of a globular domain binding site within a nucleosome. *Biochemistry* **35**, 11931–11937 (1996).
- Segers, A., Muyldermans, S. & Wyns, L. The interaction of histone H5 and its globular domain with core particles, depleted chromosomes, polynucleosomes, and a DNA decamer. *J. Biol. Chem.* **266**, 1502–1508 (1991).
- Lutter, L. C. Precise location of DNase I cutting sites on the nucleosome core determined by high resolution gel electrophoresis. *Nucleic Acids Res.* **6**, 41–56 (1979).
- Hayes, J. J. & Wolffe, A. P. Preferential and asymmetric interaction of linker histones with 5S DNA in the nucleosome. *Proc. Natl Acad. Sci. USA* **90**, 6415–6419 (1993).
- Thomas, J. O. & Wilson, C. M. Selective radiolabeling and identification of a strong nucleosome binding site on the globular domain of histone H5. *EMBO J.* **5**, 3531–3537 (1986).

- Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
- An, W., van Holde, K. & Zlatanova, J. Linker histone protection of chromosomes reconstituted on 5S rDNA from *Xenopus borealis*: a reinvestigation. *Nucleic Acids Res.* (in the press).
- Thomas, J. O., Rees, C. & Finch, J. T. Cooperative binding of the globular domains of histones H1 and H5 to DNA. *Nucleic Acids Res.* **20**, 187–194 (1992).
- Drwe, H. R. & McCall, M. J. Structural analysis of a reconstituted DNA containing three histone octamers and histone H5. *J. Mol. Biol.* **197**, 485–511 (1987).
- Staynov, D. Z. & Crane-Robinson, C. Footprinting of linker histones H5 and H1 on the nucleosome. *EMBO J.* **7**, 3685–3691 (1988).
- Pehrson, J. R. Thymine dimer formation as a probe of the path of DNA in and between nucleosomes in intact chromatin. *Proc. Natl Acad. Sci. USA* **86**, 9149–9153 (1989).
- Mirzabekov, A., Pruss, D. & Ebralidze, K. K. Chromatin superstructure dependent crosslinking with DNA of the histone H5 residues Thr1, His25 and His62. *J. Mol. Biol.* **211**, 479–491 (1989).
- Hamiche, A., Schulz, P., Ramakrishnan, V., Oudet, P. & Prunell, A. Linker histone dependent DNA structure in mononucleosomes. *J. Mol. Biol.* **257**, 30–42 (1996).
- Hayes, J. J., Pruss, D. & Wolffe, A. P. Contacts of the globular domain of histone H5 and core histones with DNA in a “chromosome” model. *Proc. Natl Acad. Sci. USA* **91**, 7817–7821 (1994).
- Howe, L. & Ausio, J. Nucleosome translational position, not histone acetylation, determines TFIIIA binding to nucleosomes *Xenopus laevis* rRNA genes. *Mol. Cell. Biol.* **18**, 1156–1162 (1998).
- Pannetta, G., Buttini, M., Flaus, A., Richmond, T. J. & Rhodes, D. Differential nucleosome positioning on *Xenopus* oocyte and somatic 5S RNA genes determines both RFIIA and H1 binding: a mechanism for selective H1 repression. *J. Mol. Biol.* (in the press).

Acknowledgements. We thank T. Hamelryck and Y. Geunens for help with figures, and L. Wyns for helpful discussions. Z.Y.B. received a grant from the VIAB. V.R. was supported by a Public Health Service grant from the NIH. This work was supported by grants from (N)FWO and the OZR-VUB.

Correspondence and requests for materials should be addressed to S.M. (e-mail:svmuyld@vub.ac.be).

correction

Weekly cycles of air pollutants, precipitation and tropical cyclones in the coastal NW Atlantic region

Randall S. Cerveny & Robert C. Balling Jr

Nature **394**, 561–563 (1998)

In the fourth paragraph from the end of this Letter, the values associated with the long and short data set differences were switched. The corrected statement should read: “For the ‘long’ data set, maximum surface winds average 3.6 m s^{-1} slower for observations made on Saturdays than those for Fridays (Fig. 2c). The unbiased ‘short’ data set has Saturday wind observations averaging 5.0 m s^{-1} slower than those on Friday; these differences are highly significant.” □

Tools of the DNA trade

The hooks and needles of molecular biology include some of the items found here – DNA and plasmid isolation kits, vectors, enzymes and labels – as well as the loom of the trade, an automated purification workstation.

HiGro

From GeneMachines

A high-capacity, oxygenated incubating shaker for bacteria and phage growth

Within its compact 56×56-cm footprint, this unit contains four growth chambers, each with independent temperature and gas flow control. The chambers hold 14 microtitre plates each, with a total capacity of 56 plates. The plates are housed in easy-to-handle cassettes that are designed to hold either standard 96-well microtitre, or deep-well plates, both of which are compatible with other GeneMachines automation components. Oxygen flow reduces carbon dioxide build-up and orbital shaking provides constant agitation for optimal growth conditions. The motion of the four growth chambers is balanced to eliminate effectively the vibration associated with many shakers.

Reader Enquiry No. 100

ProofSprinter

From Hybaid

A licensed DNA polymerase mixture for high-yield and high-fidelity PCR, up to 20 kb

The mixture combines the proofreading capacity of *Pwo* with the high process capability of *Taq*, making it a good choice for cloning or sequencing of PCR products. ProofSprinter is said to have robust performance, owing to its high DNA polymerase activity. Thus, in contrast to many other proof-reading enzymes, ProofSprinter increases PCR yields for both long and short templates, according to Hybaid. It is also said to work efficiently with templates that cause problems with *Taq* and other DNA polymerases. This mixture is available with fully optimized buffers or as a complete kit that includes dNTPs, control DNA and primers.

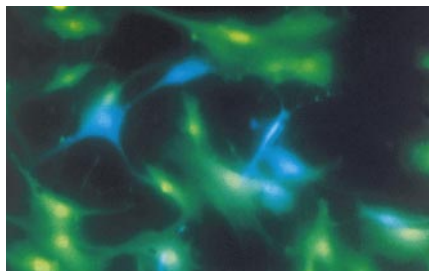
Reader Enquiry No. 101

CytoBlast

From Packard Instrument Company

A new fluorescence resonance energy transfer (FRET)-based system for the study of cis-acting sequences and trans-acting factors

This gene reporter assay system was created by Gregor Zlokarnik and Roger Tsien (at the University of California, San Diego) and relies on an enzyme-substrate reaction to monitor gene expression. A putative promoter element powers the production of β -lactamase and the level of enzyme production is detectable with the addition of the CytoBlast CCF2/AM substrate ester. Each intact substrate molecule undergoes



Have a CytoBlast, with gene expression reporting.

FRET and emits in the green spectrum when excited. β -lactamase cleaves the molecule and disrupts FRET, which shifts the emission into the blue spectrum. This emission change is directly related to the promoter activity and allows for single-cell, real-time gene expression reporting. A separate company, Omega Optical, carries a complete line of microscope filters for this assay system. CytoBlast is available to non-profit organizations only.

Reader Enquiry No. 102

PSI Clone Plasmid Prep II

From EMP Biotech

A kit that is designed for inexpensive plasmid DNA isolation

This plasmid preparation kit utilizes an anion-exchange column in a gravity/spin-column format, capturing plasmid DNA from a cleared alkaline lysate. DNA can be recovered from the eluate either by alcohol precipitation or through the use of a CentriSpin-20 spin column. The PSI Clone is said to provide plasmid DNA yields of 2 to 4 μg per ml of culture for high copy number plasmids, and 0.1 to 0.8 μg ml⁻¹ for low copy number plasmids. Plasmid DNA concentration is normally greater than 500 ng μl^{-1} for 3 to 5 ml of culture with an absorbance ratio (260 nm and 280 nm) of 1.9.

Reader Enquiry No. 103

FastTag Texas Red

From Vector Laboratories

A fluorescent nucleic acid labelling kit that uses a new coupling method to attach non-radioisotopic labels to DNA or RNA

This method uses photo- or heat-activation to link a universal FastTag linker arm uniformly to the nucleic acid. Thus, a fluorescent maleimide label is covalently attached to the nucleic acid through the linker arm. Applications for this label include *in situ* hybridization, Southern and northern blot hybridization, PCR amplification, and colony and plaque hybridization. When used for *in situ* hybridization, the fluorescent signal can be directly detected or amplified using fluorescence or enzymatic techniques. Other thiol-reactive labels can be used with the FastTag linker arm for simultaneous detection of multiple, differentially labelled probes in a single experiment.

Reader Enquiry No. 104

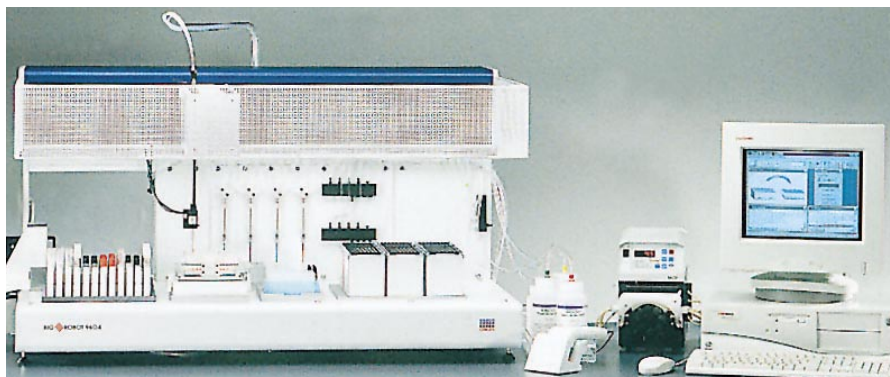
Assay tools

BioRobot 9604

From Qiagen

An automated laboratory workstation for the purification of nucleic acids

Intended for clinical and research samples in molecular diagnostic, pharmaceutical and plasma-fractionation facilities, this nucleic acid purification workstation can be used for sensitive downstream assays, such as PCR, AT-PCR, TaqMan analysis, ligase chain reaction, nucleic acid sequence-based amplification and transcription-mediated amplification. The robot is designed to provide reliable, high-throughput nucleic acid purification with ready-to-use, standardized protocols. These protocols provide automated processing for consistent yields without the use of toxic chemicals. Precise, high-speed, liquid-transfer systems help to ensure



Qiagen's BioRobot 9604 provides purification for a variety of downstream molecular biology assays.

cross-contamination-free processing for fast, reliable results. Operator interaction is reduced to increase safety for laboratory personnel. Moreover, samples are identified and process data are recorded throughout the entire run, eliminating tedious manual documentation. Data can be shared with other instruments via disk or network.

Reader Enquiry No. 105

Web-based informatics

GenWeb 4.5

From Compugen

The latest version of the company's web-based bioinformatics platform

GenWeb schedules DNA and protein database searches using BLAST, FASTA or the company's ProfileSearch and Frame+ search algorithms. The program is implemented in HTML/Java and runs on Netscape Navigator or Microsoft Internet Explorer. Designed for easy use by researchers familiar with the Internet, GenWeb provides links to NCBI's Entrez and any SRS server, as well as proprietary databases. Multiple and single results may be displayed using intuitive, colour-coded alignment maps with zoom capabilities. This program requires Compugen's GenCore operating system and can be used on any SGI, DEC or Sun UNIX workstation to provide secure data sharing over an Intranet and/or the Internet.

Reader Enquiry No. 106

On your microplate

Minifold I

From Schleicher & Schuell

An easy-to-assemble system that allows the simple application of multiple samples of DNA, RNA or protein to blotting membranes

This system has four components, with interchangeable sample application plates that allow DNA, RNA or protein samples to be blotted in a 48- or 96-well format. The sample application plates are marked numerically and alphabetically by lane and by row, and are compatible with multi-channel pipettes. An adjustable, non-corrosive metal clamping plate provides uniform pressure to reduce sample leakage from the wells. The acrylic Minifold units can be easily cleaned by washing in laboratory-grade detergent and then rinsing with 10 per cent ethanol. RNases can be removed by washing in 10



Schleicher & Schuell offer the Minifold I system.

per cent NaOH for 5 min followed by thoroughly rinsing in RNase-free deionized water. Also offered is the Delrin Dot Blot model, which can be sterilized by autoclaving. Minifold I systems can be used with a variety of vacuum sources, including a vacuum pump, a house vacuum or sink aspirator. The system comes with an incubation plate, a cutting template for assays and data sheets for accurate record keeping. Typical applications include the screening of multiple samples, gene expression studies, hybridoma screening, TCA precipitation and immunobinding studies.

Reader Enquiry No. 107

Gellin'

GeneTools

From Syngene

A stand-alone gel analysis software package

Handling all types of gel media, this program is designed for a wide range of molecular biology applications, including agarose gels, Southern blots, western blots, northern blots, autoradiography and SDS-PAGE. Other uses include optional pattern matching, databasing, colony counting, ELISA plate analysis and slot-blot work. The program features a set of analysis facilities for band quantification, band matching, molecular size determinations, quantitative PCR and volume measurements. R_f determinations, dendograms and profiling comparisons are all included as standard. According to Syngene, the package is extremely easy to use and offers many automatic routines, such as 'autotrack' finding and band matching, and yet retains the flexibility of a manual override, if required. Autotrack is said to find all the tracks on a gel and to detect distorted or 'smiley' tracks in a matter of seconds, and the band-matching facility allows the user to align bands, determine track relationships and compare track data. A full analysis automatically follows track location. GeneTools can be tailored to offer interactive control for background subtraction, profile filtering and band detection sensitivity. All results tables are configurable by the user to display exactly the required data, such as molecular weight, peak height, area, volume, band number or position. All data can be exported to other packages at the touch of a button.

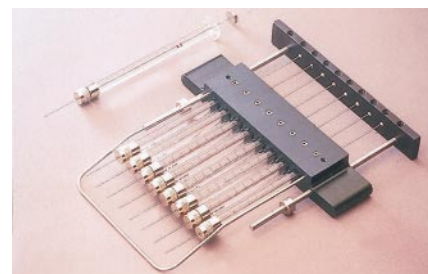
Reader Enquiry No. 108

Gel-Mate

From World Precision Instruments

Accurate gel loading of 8 or 12 samples into square or shark's tooth DNA sequencing gels

Different needle diameters allow easy loading of sequencing gel wells ranging from 0.2 mm to 0.4 mm. A needle guard prevents damage needle during sample loading, and replaceable stainless/fused-silica needles



Gel-Mate, WPI's sequencing gel-loading system.

and single syringes are available.

Reader Enquiry No. 109

The end of the tail

Mouse tail extraction

From AutoGen

Automated genomic DNA extraction from murine tail tissue

The process is said to yield high-quality DNA automatically, in yields that are comparable to manual extraction. A gentle action is said to minimize shearing, thus producing higher molecular weight DNA. The DNA recovered from the AutoGen system is ready for use in such techniques as PCR, cloning and restriction digests with no further treatment.

Reader Enquiry No. 110

These notes are compiled by Brendan Horton from information provided by the manufacturers. For more details, fill in the reader service card bound inside the journal.

correction

In *Nature* (394, 916), "a range of receptors, receptor agonists, antibodies and inhibitors" was attributed to Research Biochemicals and Semat. However, these products are offered by Semat only, and not by Research Biochemicals.

ADVERTISEMENT

PIERCE

PowerSignal™ ELISA Chemiluminescent Substrate*

Femtogram-level Sensitivity and Low-level Linearity for Luminometer-based ELISAs and HTS Applications

Features/Benefits:

- Delivers rapid kinetic light output and high signal intensity
- Exhibits a very linear signal conducive to low-level detection and quantitation
- Provides exceptional stability — reagent shelf life is at least six months at room temperature

Tel: 800-874-3723 or 815-968-0747 • Fax: 815-968-7316
CS@piercenet.com • Internet: www.piercenet.com

* Patent Pending

READER ENQUIRY NO. 48

Taking knowledge from bench to bank

Academic research provides the basis for many commercial ventures; investors are confident that they will go on reaping the rewards. But it must be remembered that scientific knowledge is the property of all, not just the few.

Brendan Horton

A recent policy statement published by the US Committee for Economic Development (CED) reports that two important trends are helping to reshape university research. One is the ability of universities and academic researchers to reap financial rewards from inventions licensed by industry. The other is the growing number of industry–university collaborations.

CED, an industry-backed think-tank, argues that government support for basic research is critically important to US industry, but that industry should pay for the development of technology. The report, *America's Basic Research*, explains that because university inventions often come about through the use of public funds, these institutions are obliged to further their use in society. This responsibility goes beyond the need to freely disseminate the descriptions of inventions, extending to how inventions are commercialized as patents (implying a restriction on use).

Contracting with society and industry

Balancing the benefit to society while providing an incentive to a commercial partner can be done using exclusive or non-exclusive licences. The development of very early-stage platform technology, which entails substantial risk, might justify an exclusive licence from a university's technology licensing office. In these cases, CED explains, the protections of both the patent and the exclusive licence are essential for a company to invest time and money in a new technology.

On the other hand, a research tool potentially benefiting the development of a range of new products in multiple companies is typically commercialized through a non-exclusive licence made widely available on reasonable terms, such as a minimal annual fee on starting and a small royalty on sales. One shining example is the \$38.5 million generated last year for Stanford University and the University of California, San Francisco, from gene-splicing technology.

The CED reports that direct industry-sponsored research at universities, though small in relation to government funding, has expanded from 4 per cent of total support for academic research in 1980 to 7 per cent in 1996. Separate from this was the passage in 1980 of the Bayh–Dole Act, which allowed researchers to retain title to inventions developed with government funding. According to Lita Nelsen, director of the technology licensing office at the Massachusetts Institute of Technology (MIT), universities have the legal right to retain royalties from such licensing and can give a fraction of the royalties as personal income to inventors. "By law, the university's share of the royalties must be ploughed back into its research and educational activities," says Nelsen.

Driving the increase in collaborations from the academic side, says Nelsen, is the feeling that the universities need to show that federal support of fundamental research results in economic growth, product development and job creation. "Technology transfer is a way of continuing our public mission and maintaining public support for universities." Also contributing to the

increase, she says, has been the incorrect prediction that government funding of basic research would be greatly diminished after the end of the Cold War.

On the industrial side, the reasons are very different, says Nelsen. "Technology is moving so fast, no one company can do all the research that is needed to keep new products in their pipeline." Further, she says, because of serious downsizing, companies are less able to develop their own technology and need to outsource research.

Managing the risks

According to David Allen, director of the office for technology licensing at Ohio State University (OSU), technology offices typically receive invention disclosures from faculty. They secure these as intellectual property, mostly through patents, and then seek licences to commercialize the patents. OSU is committed to providing value to the technology sector, says Allen. "With platform technologies, most large companies are not interested because they typically want to buy a finished product, with all the risk wrung out of it." Innovative technology development companies or entrepreneurs exist to take the newest technology forward as they know how to manage this process, he says.

The CED report stresses the importance of the university's traditional role as a source of new information gleaned from basic research and its obligation to disseminate new research findings freely. "In cases where new knowledge has commercial potential, however, patenting and licensing may be appropriate and in the public interest," but this should not compromise the educational and research missions of the university, says the report.

Around 73 per cent of research publications referenced in industrial patents were derived from government-funded research. A gauge of the impact of Bayh–Dole is the number of patents issued to academic institutions, which increased from 249 in 1974 to 1,761 in 1994. Much of this improvement in the technology-transfer process, according to Nelsen, has occurred because of the clarification of a university's right to own the patents arising from federally funded research.

Technology transfer is maturing: the result, says Nelsen, is greater uniformity in the policies that govern sponsored research at major universities. Looking back, she says, "you see a big difference in the late 80s. The learning curve accelerated as then people

Table Useful sources of further information on the Internet

UC BioSTAR Project biostar@uclink4.berkeley.edu http://www.biotech.berkeley.edu/	EBI Industry Support http://industry.ebi.ac.uk/
UCSD CONNECT http://darwin1.ucsd.edu:8000/	Programme Scripps Institute, Research & Faculty http://www.scripps.edu/research/
AUTM survey (includes data from 89% of the top US research universities) http://autm.rice.edu/autm	Novartis Biotechnology http://www.novartis.com/biotech/index.html
CHI/KPMG Peat Marwick report California: The Biomedical Frontier 1998 Report on the Biomedical R&D Industry http://www.chi.org	Lion Bioscience AG http://www.lion-ag.de/
MIT's technical licensing office http://web.mit.edu/tlo/ www/	NSF science & engineering indicators 1998 http://www.nsf.gov/sbe/srs/seind98/start.htm (http://www.york.ac.uk/org/auril/aurilink.htm)
Committee for Economic development http://www.ced.org	Includes links to related organizations, for example industrial liaison offices in UK universities.
Medicon Valley Academy http://www.mva.org	For access to one of the newsletters about the UK's department of Trade and Industry's link programme: http://www.dti.gov.uk/ost/link/news3.htm
Copenhagen Capacity http://www.copcap.com	Information about a European conference on academic industrial collaboration can be found at: http://cordis.lu.uk/home.html
Structural Bioinformatics, Inc. http://www.strubix.com	For OECD's perspective on science and innovation: http://www.oecd.org/dsti/sti/s_t/inte/index.htm
	University-related research–technology parks http://davistech.com/test/why_casestudies.html

didn't know what they were doing and expectations were unrealistic."

To take an example, since MIT reorganized its office of technology transfer in 1985 the income of the office has increased more than fivefold and the staff has increased about fourfold. The income received in 1997 was a record \$21.2 million, up from \$2 million or so in 1985. Overall, MIT consummated 59 new licences (now totalling 455 active licensees), and eight new companies were started in 1997.

Further, about 115 start-up companies have been created by MIT's technology licensing office, now one of the top technology transfer institutions. According to CED, there are more than 1,000 MIT-related companies in Massachusetts, with about 125,000 workers in Massachusetts alone and worldwide sales of more than \$53 billion. Similar developments have taken place in California's Silicon Valley and in Research Triangle, North Carolina.

Recent makes and models

Vehicles for commercialization are coming in different shapes and sizes. With institutions like MIT and Stanford setting the precedent, others would surely like to emulate their success. OSU, for example, has

adopted new mechanisms to develop more efficient technology transfer, and to create some of the entrepreneurial, business and financial infrastructure found in places like Boston or San Francisco's bay area.

Balanced properly, such endeavours should benefit the companies that use university research, the university itself, its students and the local economy. "We are more likely to evolve relationships with them to do sponsored research, experiential learning opportunities and internships for students, as part of a whole realm of activities that are partnership oriented," says Allen.

Because the university's first research park was not successful, OSU set up a 40-member task force four years ago to try to make technology transfer more effective. One of the problems, says Allen, "is that as a unit the university was imbued with all of the consensus decision-making and deliberation that occurs at a pretty glacial pace for technology transfer". The task force has created a corporation outside the university to produce entrepreneurial activity quickly and help build the infrastructure which is underdeveloped in the Columbus area. This result is the Science Technology Campus Corp., which houses 14 companies, half of which owe their existence to university technology.

Also separate from the university is the business technology centre where prospective businesses can come to have their ideas developed. According to Allen, start-up companies must be based on a relatively new technology that has a good market trajectory — that is, it must be more than just a research finding. Also, the company should have some of its leadership in place.

One of the things that the business technology centre offers is a really strong board of trustees; these are entrepreneurs, business people and financial people, says Allen. Success depends on effective teamwork with the board, he adds.

Faculty members seldom leave their positions at the university, but often stay involved on the research side "by pushing the technology in the laboratory forward", he says. However, they often have colleagues, friends or graduate students who want to capitalize on the technology through a corporate venture. The business technology centre then adds to this initial team by filling out the business structure. Funding comes from risk capital sources, often 'business angels' — wealthy friends or associates of entrepreneurs. Allen would like this source of capital to be organized into a formalized mechanism. □

Partnerships and the critical mass

Academic faculties today may have insufficient openings to accommodate all prospective candidates, but industry partnerships are generating opportunities for basic and applied researchers to help make up the difference. Funding indicators reported by the US National Science Foundation confirm that collaborations between the academic and industrial sectors are on the rise. In the United States, investment by industry in basic research performed at universities and colleges increased in real terms between 1991 and 1997, rising to a total of \$1.05 billion. An even longer term trend is indicated by the fact that co-authorship of journal articles by US industrial researchers with their colleagues in academic and government labs has been increasing steadily across all fields since 1981.

The trend towards industry collaborations has been most pronounced in the United States, thanks to the great depth of its scientific establishment, and a well-developed infrastructure of venture capital and technology transfer. But, as the worldwide force of scientific creativity strains to find expression, industrial-academic partnerships, either on the US or some comparable model, are flourishing elsewhere. In Europe, for example, the member states of the European Molecular Biology Laboratory (EMBL) have authorized the lab to accept

shares in companies in exchange for releasing technology, and to establish a technology-transfer company fully owned by EMBL.

The European Bioinformatics Institute (EBI), an outstation of EMBL at Hinxton Hall, in Cambridge, has an active outreach programme to foster company-funded research. This has led to important advances in development of CORBA, a set of standards which allows databases and software applications to communicate. Once developed by the company/institute consortium, the standards can be used to develop fully commercial products.

One collaboration between EMBL and the private sector is Lion Biosciences, based in Heidelberg, Germany, which is focused on high-throughput expression profiling and bioinformatics. Supported by private and state funding, it has formed relationships not only with EMBL/EBI, but also with the German Cancer Research Centre, the University of Heidelberg, Hoechst, and other companies. Since its foundation 15 months ago, Lion has doubled its staff numbers to 55, and expects to double again within a year.

Christian Marcazzo, a product development manager of Lion's sequence retrieval system, an indexing and retrieval tool for molecular biology data libraries, which was developed at EMBL/EBI over the past six years or so, says he expects hiring in



Marcazzo: expects hiring to triple.

his department to triple in the year ahead. The new staff will include not only scientists but people who are not usually in strictly academic development projects: quality assurance, documentation, customer support and marketing. "The goal of the activity," he says, "is to turn good

technology into good products for the pharmaceutical, biotech and agriculture research communities."

Marcazzo, who recently moved to Lion from a job in California's Silicon Valley, observes that the Europeans are eager to explore collaborations and to commercialize academic research. Nevertheless, the road for start-ups is not as well paved in Europe as in the United States. "There are fewer venture capitalists, and they are more conservative. The consultants you can find practically on every street corner in Silicon Valley are not here. And going public in Europe involves a lot more than merely satisfying minimum capital requirements to get a NASDAQ Stock Exchange listing, as in the States." He adds that employee stock options are unusual in

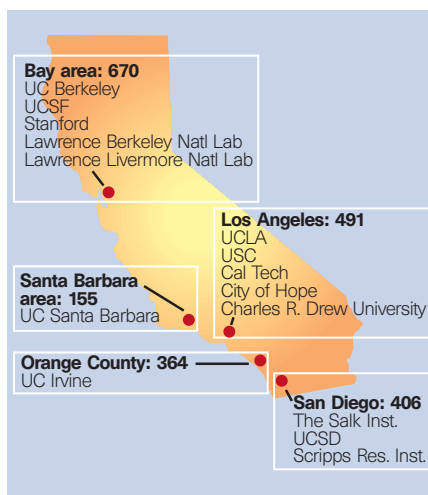
How to turn R into D and get paid for it

Diane Gershon

It is hard to find reliable and useful data on the economic benefits of academic research. One study by Edwin Mansfield¹ calculated the social rate of return from academic research during 1975–78 to be 28 per cent. Mansfield cautioned, however, that his estimate was “at best a very crude beginning”. The Association of University Technology Managers (AUTM; see table) more recently estimated that, in 1996, technology transfer from academic institutions to industry contributed more than \$24 billion to the US economy and supported about 212,500 primarily high-skilled, high-paying jobs. Comparable figures for the previous year were \$21 billion and 180,000 jobs.

Although clearly not every licensed invention leads to a marketable product, much less a lucrative one, many have resulted in significant new products and technologies. Some have even formed the basis for new companies or entire industries.

Biotechnology is a good example of an area where public investment in basic



Hotbed clusters of medical technology activity in California: companies are concentrated around major research institutions.

research and evolution of the industry are inextricably linked. Small, entrepreneurial businesses in biotechnology and other areas of high technology were the engine for eco-

nomic growth in California, which led to the dramatic turnaround in the state's economy.

Healthcare technology as a whole, of which biotechnology is a significant part, is now one of the fastest-growing, highest-paid industries in the state, second only to electronics in the high-technology arena, according to a report issued in July by the California Healthcare Institute and KPMG Peat Marwick (see table). In 1997, more than 200,000 people were employed in healthcare-related jobs in the state, with average annual earnings of \$50,500 (compared with state-wide average of \$32,800 for all industries). The highest-paid positions in the sector were in biotechnology, where the average salary was \$67,000.

Perhaps the most compelling evidence of the vital link between academic institutions and the biotech industry is reflected in the results of a recent survey of biotech companies carried out by the University of California's (UC's) critical linkages project. It showed that one in three US biotech companies is within 35 miles of a UC campus; that one in five of California's biotech companies

DATA FROM CHU/KPMG REPORT

Europe, and the start-up culture of “high-risk, high reward” is not a familiar idea to most young Europeans.

France, too, has tended to lag behind the United States in developing contacts between the academic community and industry. Daniel Louvard, research director of the Institut Curie in Paris, says: “A certain bias against applied research has existed in France. But this attitude is changing rapidly, driven in part by budgetary pressures.” Already there has been an increase in the number of contracts between industry and the Curie, and ten patents have been granted so far this year to Curie researchers.

Elsewhere in the world the atmosphere may be less conducive for collaborative research ventures. Fred Dotzler, an investor with Medicus Ventures in Palo Alto, California, says: “We were recently involved with a venture in Korea and I got the feeling there was no vehicle whatever over there for technology transfer from the academic sector to start-up companies.” But once a collaboration is in place its benefits — an accelerated pace of discovery and generation of science jobs and revenue — accrue not only to small ventures with rapidly evolving new technologies, but to mature, diversified establishments as well.

An example of the latter sort of partnership is the one between Novartis, the

Swiss pharmaceutical giant, and the Scripps Institute, the graduate research university in La Jolla, California. Beginning last year, Scripps is receiving \$20 million a year for five years from Novartis, with an option for a five-year renewal in 2002. The form of the funding is of particular value to Scripps, because 65 per cent of it is “unrestricted”, meaning it can be used to pay for physical plant and equipment, expenditure that is excluded from the \$70 million a year that Scripps receives from the US National Institutes of Health (NIH). Some \$7 million a year of the Novartis funding is project-specific, and is allocated among the institute's labs by the Scripps scientific advisory board, on which Novartis has non-controlling representation.

Does the presence of industry money in a partnership change how science is done? Certainly the application of resources to a research venture advances the possibility of discovery and extends avenues of enquiry. But can the funding also — especially when concentrated and of long duration, as in the Scripps/Novartis partnership — exert some unwholesome influence on the conduct of science? Not in the least, asserts Arnold LaGuardia, Scripps senior vice-president. “We're basic researchers and we don't change over to applied research just because industry money comes in the door.” To the contrary, he sees a salutary influence beyond the dollars, in

that exposure to the industrial environment can give academic researchers a heightened appreciation of the difficulties faced by their colleagues in industry and applied science.

Nevertheless, it is pertinent to remember that a proposed agreement in the early 1990s between Scripps and Sandoz (a predecessor company of Novartis) caused a furore because of provisions that appeared to limit the autonomy of Scripps and its researchers. Objections by Congress and NIH officials caused that agreement to be renegotiated, giving rise to the one in place today.

The job-seeking researcher dismayed at the prospect of science defined by state and corporate policy-makers, subject to the stipulations of patent lawyers and venture capitalists, can be comforted by remembering that it is the individual idea that lies at the heart of science. Ernest Beutler, chair of molecular and experimental medicine at Scripps, says that the idea that large teams — the so-called ‘critical mass’ — are the key to success in science must be rejected. “Collaborations do play a role in exploiting new ideas, but the origin of an important new concept can usually be attributed to one person. As the Nobel laureate Charles Huggins once said, ‘the critical mass in science is one scientist’.” **Potter Wickware** *Potter Wickware is a science writer in Oakland, California, USA. e-mail: wick@netcom.com*

was founded by UC scientists; and that six of the top ten best-selling biotech drugs in the United States stemmed from UC research.

The role of the state's research enterprise in the regional economy has long been recognized by Richard C. Atkinson, former chancellor of the University of California at San Diego (UCSD; see below) and, since 1996, president of the UC system, which includes nine academic campuses. Atkinson's priorities have been to make new relationships with the business community to boost economic growth, to create high-wage jobs, and to enhance California's competitiveness. He has created a series of industry-university cooperative research initiatives in areas of the state's economy which are thought likely to benefit the most from basic research. These are biotechnology (the BioSTAR project), multimedia and semiconductor manufacturing, soon to include informatics and communications.

BioSTAR: a three-way partnership

BioSTAR (see table), a matching grant programme that partners university scientists with California biotech businesses, aims to draw companies into the earliest stages of research, when they are most likely to pick up something truly innovative.

In the first two years since the programme has been operational, says director Susanne L. Huttner, BioSTAR has invested \$22 million in new research partnerships. Grants run for two years and average about \$250,000, although participants are eligible for a competitive renewal process. About 300 UC faculty and graduate students are currently in the programme, which provides graduate students with a sense of the relevance of their work and an opportunity to see how businesses approach research early on in their careers, Huttner says.

Even though BioSTAR is in its infancy, Huttner already has anecdotal evidence that some participating companies have been able to attract more outside investment by virtue of their alignment with a research institution. She says investors view this as a means for companies to expand their R&D capacity while buffering risk, as they are contributing only half the costs. It also provides companies with a fairly clear path to protecting their intellectual property rights.

Encouraging entrepreneurship

But as David L. Gollaher, president and chief executive of the California Healthcare Institute (CHI), writes in the foreword to the CHI/KPMG Peat Marwick report: "robust science alone does not build companies. The secret of California's success... has been the emergence of a vibrant market in company formation". Access to capital through venture capitalists and investment bankers, as well as to financial consultants and other service providers, all play a part in creating an

environment that embraces and encourages entrepreneurship.

Organizations like industry-backed UCSD CONNECT (see table), which currently serves the San Diego region, are acutely aware of this. It is no coincidence that UCSD CONNECT was founded 12 years ago, when the university had lost a major semiconductor research project to Austin, Texas — in part, because the entire community in Austin rallied behind the university. This, coupled with the loss of some 10,000 jobs a few years later when General Dynamics closed shop and moved to Tucson, Arizona, caused the San Diego economy to go belly-up. With local initiatives like UCSD CONNECT, the recent growth in the number of high-tech jobs in the region has now more than replaced earlier losses. In fact, the telecommunications firm Qualcomm, which was co-founded by a faculty member from UCSD and one from the University of California at Los Angeles, has now almost by itself replaced General Dynamics.

Although not in the technology-transfer business itself, UCSD CONNECT helps economic development in the region by offering a range of 'networking' programmes for high-tech companies in areas such as biotechnology, computers and telecommunications. However, companies don't need to

have ties to UCSD to access its services. "Our uniqueness is our emphasis on entrepreneurship," says Bill Otterson, director of UCSD CONNECT. Over the past 10 years, its efforts have resulted in almost \$5 billion in deals with participating companies.

UCSD CONNECT's Springboard programme, for example, takes a group of people with an idea for a company and teams them up with individuals who are knowledgeable in the field but not competitors. These advisers provide their time and expertise *pro bono*, coaching groups on business plans before presentation to potential investors. Even 'virtual boards' can be set up in the short term for companies that can't afford an actual board.

"The first need is to get people with a strong scientific bent to think in terms of marketing," says Otterson. He came to UCSD not as an academic but as an entrepreneur with a business background, and says everybody is looking for the next Bill Gates. Although San Diego doesn't have a Gates, it is now a leader in wireless communications, which came out of UCSD, and is one of the top biotechnology areas in the country. □

Diane Gershon is assistant editor, new technology, at Nature Medicine.

1. Mansfield, E. Academic Research and Industrial Innovation. *Res. Policy* 20, 1–12 (1991).

Medicon Valley: a bridge to collaboration

Medicon Valley is hoping to follow the successful lead of the technology boom towns. Deriving its name from Silicon Valley, this region owes its existence as an economic zone to the building of a bridge across the Øresund Sound — scheduled to open in 2000 — which will link Copenhagen in Denmark to the city of Malmö in the south of Sweden. The intent is to capitalize on the region's strength in healthcare, pharmaceuticals, biotechnology and life science research, says Jes Østergaard, managing director of the Medicon Valley Academy.

The platform of many technological regions is the strength and abundance of university research. Medicon Valley has Lund University (Sweden's largest) and the University of Copenhagen, as well as the Technical University of Denmark, the Copenhagen Business School of Pharmacy and Malmö University. Moreover, 50–60 per cent of the pharmaceutical and biotech companies in Scandinavia reside in the valley, as well as a fair number of the medical device companies, according to Medicon Valley literature (see table).

The region offers clinical-trial services and medical collaborations through its large university hospitals and other regional hospitals. There are also private and government research institutes, such as the

John F. Kennedy Institute for neurology and genetics, the Hagedorn Institute for Diabetes and Molecular Endocrinology, the Wallenberg Neuroscience Centre (with a new gene therapy division) and the Danish Cancer Society's Cancer Epidemiology Centre.

The academy

The Medicon Valley Academy was created to further the relationship between the universities, industry, hospitals and life-science-based companies in the area. It was established more than a year ago by six of the area's major universities with offices on either side of the sound in Copenhagen and Lund. As Østergaard puts it: "We assist with the development of collaboration by adding infrastructure and networking with projects. And we are building a database of information about research and knowledge in the area and making it visible and more accessible to those in the region and those outside."

Operating across the Danish-Swedish border has the added benefit that the academy can draw on money from local universities, which have contributed \$1.5 million, and from the European Union, which has contributed \$750,000. This money, while modest, is meant to catalyse the formation of the network, says Østergaard. The academy

From research to commercial benefits

Helen Gavaghan

Somehow it seems appropriate that John Daugman's first degree is in physics and philosophy. Now in the computing department at the University of Cambridge, he has in the past formed two companies, one hardware and one software, thus reconciling academic research and commerce.

With a valedictory shipment of neuroscientific instruments to the Amazon rainforest, Daugman is winding up his hardware company, a small enterprise that sprang from research into visual neuroscience and which he founded in 1984 while at Harvard. His second company, IriScan, opened in 1993. It licenses Daugman's algorithms for iris recognition which, because the iris remains unchanged throughout life and is unique to each individual, may replace existing ways of identifying people. The system springs from Daugman's basic research into neural computing, and investment bankers J P Morgan estimate the market value of the technology is now \$100 million.

Much as governments would like more university researchers to found companies

that create jobs, entrepreneurial academics are still few and far between, according to Jim Reed from the University of Surrey — "perhaps no more than 2 per cent in the UK," he says. The picture is similar in Europe, although US academics have traditionally shown more zeal in this regard. Nevertheless, Europe's researchers seem likely to be encouraged along the same path as their transatlantic colleagues. In its 1998 report, *Science, Technology and Industry Outlook*, for example, the Organization for Economic Cooperation and Development (OECD) foresees no end to the fiscal pressures on the pot of public funds for research.

One of the first academic institutions to feel the bite of public funding cuts and successfully seek industrial collaboration and spin-off companies was Trinity College, Dublin. "It was necessary to survive," says John Hegarty, dean of research at Trinity. The university's entrepreneurial academics are able to stay with their departments, so keeping knowledge flowing between the two sectors, while receiving help with intellectual property issues or marketing from the university's innovation centre when setting up

their companies. One company, Iona Technology, which provides software to enable different computer applications to work together, now employs 500 people. "It set the tone for entrepreneurship in Ireland," says Hegarty, "and it was partly because of Iona that the government began to see the role of universities in economic development."

The fiscal pressures highlighted by the OECD, coupled with the potential that high-tech innovative companies have for job and wealth creation, are pushing governments to formulate policies to encourage ever-closer links between industry and academic institutions. Such cooperation extends beyond spin-offs to collaborations on either specific projects or strategic research as well as training, testing and validation of technical ideas. From industry's perspective, too, collaboration with academics makes sense. Tougher competition among companies developing fast-moving technologies is forcing them to turn research into products ever more rapidly. Although larger companies may survive in these turbulent waters, small and medium-sized enterprises often do not have the resources to keep the product line full.

does not pay people to carry out projects, he says: "We bring people together who have interest in the end goal and facilitate a project's completion."

Valley vistas

Two new research centres are opening in the valley with investments totalling \$125 million each, according to the academy. Already open is a biomedicine centre at Lund University. The biotechnology centre at the University of Copenhagen has yet to break ground but has been designated funds for 1999. Both are structured to facilitate collaboration in research.

The centre in Copenhagen will be located at the main university hospital, says Østergaard, and will house a collection of laboratories and offices for small- and medium-sized biotech companies, university researchers and hospital staff.

Financing is obviously an important part of the equation for small start-ups and companies building on licensed or proprietary technology. There are many privately held business funds and financing institutions in the Medicon Valley (see table). There is public debt financing through low-interest loans, or business development finance, which provides a low-risk environment in which the loans are only paid



Øresund bridge: gateway to the growth of new companies in two countries.

back if the venture is successful. In the event that the venture is abandoned, the loan is exchanged for the rights to the project, and business development finance simply writes it off. This public funding is also available for foreign companies doing business in the region. According to the academy, business development finance has provided capital for 110 projects in the medical sector, guaranteeing up to 45 per cent of the total development cost to a maximum of \$7 million.

High capacity

For San Diego-based Structural Bioinformatics, the presence of Copenhagen Capacity was a big reason for opening a

subsidiary in Hørsholm, adjacent to the Technical University of Denmark. Copenhagen Capacity aims to provide services and information to foreign companies, considering greater Copenhagen as a possible location for activities. According to Susan Burgess, Structural Bioinformatics' vice-president of corporate development, Copenhagen Capacity helped by introducing the company to investors and providing guidance and counselling. "Our last round of investment was largely through several Danish funds. We also got debt financing through the Danish government," she says.

Structural Bioinformatics' Danish subsidiary now has six full-time employees, and works with three faculty members from the Technical University of Denmark, who act as scientific advisers to the parent company and as strategic and technical directors to the subsidiary. It was Copenhagen Capacity that introduced Structural Bioinformatics to Jakob Bohr, one of the Danish scientists with whom they are now working. As it happened, Bohr and his brother had some interesting technology that was applicable to protein folding. This was a good fit with the company's core technology, which is based on three-dimensional structure modelling and protein structure-based algorithms used in drug discovery.

B. H.

Yet it is precisely these smaller enterprises that the OECD and national governments have identified as showing the greatest potential for creating jobs. In the face of this reality, Europe's governments are paying particular attention to developing policies supportive of this sector. The European Union's largest scheme encouraging collaboration — the fifth Framework programme — will, for example, set up a special office to help small companies interested in collaboration to find partners. The programme will run from 1998 to 2002 and spend a little over £9 billion (US\$15 billion).

Although the formal requirement for receiving Framework money (up to 50 per cent of a project) is that two organizations from different countries collaborate with each other, the practice is a little different, says David Moore of the UK's Office of Science and Technology. Proposals with various partners, including some from industry, particularly smaller companies, stand a better chance of winning support, he says.

Each European country, too, seems to be directing its national research and industrial policy towards encouraging greater collaboration with small and medium-sized companies. A key player in Denmark, for example, is the Riso National Laboratory. In the past, industrial-academic collaboration tended to focus on one specific project. Now strategic collaboration, in which the aim is to keep the product pipeline full, is important, says Jorgen Kjems, Riso's managing director. Given that Danish industry includes many medium-sized, high-tech companies that operate in highly competitive global markets, Riso's approach is important.

Other Danish government schemes include a four-year programme to stimulate start-ups. Grants of up to \$116,000 are available to help young scientists formulate business plans. Riso has also launched several companies. Some took off immediately but for others, says Kjems, "we were a patient nest". Ibsen Microstructures, for example, took five years to become established.

Cultural barrier

A significant question for academic-industrial collaboration both within and between countries is whether the partners can understand one another. The language and cultural ethos of academic science is clearly different, at least in stereotype, from that of business. The traditional view, says Alan Brickwood, a professor in the design research centre at Britain's Brunel University, is that the two have a bruising encounter, exchange money and knowledge, then rapidly go their separate ways. After research undertaken on secondment to the UK Department of Trade and Industry (DTI), Brickwood says, "matters are not so simple". The DTI is currently looking at what motivates industry and academics. An interim report released in July

gives the academics' viewpoint¹.

Brickwood's team invited the 250 UK institutions of higher education to give examples of what they considered to be successful collaborations with industry. Of these, 89 responded, giving 250 examples. Brickwood concludes that, for successful collaboration, academic institutions do not need to change their values, but will have to change their behaviour. As expected, academics value curiosity-driven research and see benefit in the money from industry. Yet Brickwood suspects that industry will be surprised to learn that many academics see more benefit in their strategic discussions with industry than in the financial rewards.

A few anecdotal comments to *Nature* suggest this phenomenon may be reasonably widespread across Europe. "Collaboration with industry keeps us aware of actual industrial developments," says Pierre Mathys, an engineering professor at the French Free University of Brussels. One example of demonstrating the potential for shared values and mutual benefit springs from the work of Andy Brass, a professor of computing who runs the bioinformatics postgraduate programmes at the University of Manchester in the United Kingdom. After a trade mission to the functional genomics companies on the west coast of the United States last February, Brass returned to Manchester with strengthened contacts.

He was energized by his tour, and, in his own words, he was feeling "gung ho". With help from the university's unit for supporting start-up companies, Brass set up a company, Bioinformatics Solutions, offering the information technology skills that biotech and genomics companies need to sift through mountains of data in search of the crucial information to make their fortunes. Brass retains copyright on his algorithms, while the data clearly belong to the company for which he works. Yet financial reward is not his sole driving force. "Working with the small/medium enterprises rather than big pharmaceutical companies gives you a sense that you are really contributing something," says Brass. And the icing on the cake is that "the companies have a lot of big data sets, and I get to play with other people's data".

Which company owns which piece of intellectual property may well be clear for Bioinformatics Solutions and the biotech companies it will work with, but collaborations cannot always rely on such logical separations. The conflict between the academic's need to publish freely and the commercial need not to give away information undermining patent claims also remains.

Jan Chojceki, managing director of the Norwich-based technology-transfer company Plant Biotechnology, reviews the issues in *Getting the best out of academic intellectual property*, published in the October 1997 issue of the EU-sponsored newsletter *PIP*

(*Plant Industrial Platform*). Identifying intellectual property that the inventors have not already disclosed is, he writes, a problem. "Speculative comments in one paper about what the investigators would go on to do next can lead to patents claims on subsequent intellectual property being turned down on the grounds of obviousness." The other side of this coin is that the academic idea might be so far removed from proof of concept that it could not get a patent.

If industry is to collaborate then both issues have to be addressed. When there are several collaborators, the issues, not surprisingly, are more complex. Solutions such as granting licence options during a period of evaluation and working out formulas for assigning economic value to the various rights need careful thought, says Chojceki.

Financial institutions

The investment community is increasingly recognized as having an important role in developing innovative ideas from a country's science base. If industry and academics speak a different language and have a different ethos, bankers have yet another. There is "an empathy gap", says Graeme Jones, head of the Innovation and Growth Unit at NatWest bank in the United Kingdom. Innovative businesses often have the potential for very fast growth and need to compete immediately in a global marketplace, requiring considerable business skill. Yet the originators of the idea often do not have these skills. The combination of projections of rapid growth, cutting-edge technology and limited business and marketing skills makes bankers nervous, says Jones.

In an effort to bridge that gap, NatWest bank established a service in 1989 in which it keeps aware of government schemes and grants and runs a network of 'business angels'. Deutsche Bank, ING of the Netherlands and NatWest are to evaluate whether this new-technology appraisal service could be applied to loan decisions for small high-tech companies throughout Europe.

Adrian Hill, head of the University of Bristol's research support industrial liaison organization, discerns and welcomes a new ethos that makes it respectable to move knowledge out of the universities via the commercial sector, but he hopes the transition to purely commercial ways of thinking will never be fully achieved. "That could only be accomplished by selling out completely," he says. Such a sell-out would profoundly undermine the science and technology base that schemes encouraging collaboration between industry and academic science are keen to exploit. □

Helen Gavaghan is a freelance science writer based in Hebden Bridge, Yorkshire, UK.
e-mail: 101543.30@compuserve.com

1. Brickwood, A. *Higher Education Working with Business*. Interim report (DTI, London, 1998).